RPZ exercise book

October 21, 2025

Generated from cbf36bbd at https://gitlab.fel.cvut.cz/rpz/exercise-book

Contents

Basic Probability **3**
- Conditional Probabilities . . . 3
- Cancer Test Problem . . . 4

Bayesian Decision Making **5**
- Umbrella Rain . . . 5
- Coarse Decision Space . . . 6
- Gaussian Three Classes . . . 7
- Bayes Classification Gaussians . . . 8
- Gaussian Conditionals . . . 9
- Error Correcting Codes . . . 10

Non-Bayesian Decision Making **11**
- Minimax with Discrete Measurements . . . 11
- Neyman-Pearson with Discrete Measurements . . . 12
- Minimax - Continuous Measurements . . . 13
- Neyman-Pearson: Continuous Measurements . . . 14

Non-Parametric Density Estimation **15**
- Parzen Window Density Estimation . . . 15
- k-NN Classification . . . 16

Support Vector Machines **17**
- Hard Margin SVM . . . 17
- Soft Margin SVM . . . 18
- Graphical Interpretation . . . 19
- Dual Task . . . 20
- Kernels . . . 21

Neural Networks **22**

Basic concepts 22
Computational graph 23

Solutions **24**

Basic Probability

Conditional Probabilities

Consider the same example as in the lecture. The joint probability $p(x, k)$ is given by the table:

	x=1	x=2	x=3	x=4
k=**rain**	0.02	0.12	0.09	0.04
k=**no rain**	0.38	0.28	0.06	0.01

where $k \in \{\text{rain}, \text{no rain}\}$ and $x \in \{1, 2, 3, 4\}$ denotes the level of cloudiness.

1. Compute the marginal probabilities $p(k)$ for all values of k and $p(x)$ for all values of x.
2. Compute the probability that the cloudiness is less or equal than 2 given that there was rain.

↓ View the solution

Cancer Test Problem

Suppose we have a test for cancer with the following statistics:

- The test was positive in 98% of cases when subjects had cancer.
- The test was negative in 97% of cases when subjects did not have cancer.
- Suppose that 0.1% of the entire population have this disease.

A patient takes a test. Denote the variable that the patient has cancer as $C \in \{y, n\}$, and that the test is positive as $T \in \{+, -\}$.

1. Compute the probability that a person who tests positive has this disease.
2. Compute the probability that a person who tests negative does not have this disease.

↓ View the solution

Bayesian Decision Making

Umbrella Rain

Consider the joint probability $p(x, k)$ given by the table

	x=1	x=2	x=3	x=4
k=**rain**	0.02	0.12	0.09	0.04
k=**no rain**	0.38	0.28	0.06	0.01

where $x \in \{1, 2, 3, 4\}$ denotes the level of cloudiness.

You have three possible decisions $D = \{\text{umbrella}, \text{no umbrella}, 100\}$ to make on a given day:

- umbrella : you take an umbrella with you,
- no umbrella: you do not take an umbrella with you and if it rains, you will get wet,
- 100: you do not take an umbrella with you but you make a fixed decision that if it rains, you will buy a new umbrella for 100 CZK.

Let the loss (cost) matrix $W(k, d)$ be as follows:

	umbrella	no umbrella	100
rain	0	10	5
no rain	5	-2	0

Compute:

1. The chance of rain given the cloudiness 2
2. The expected cloudiness on a rainy day
3. The risk of not having umbrella if the cloudiness is 2 (called partial risk)
4. The risk of not having umbrella ever
5. The risk of always carrying an umbrella
6. The optimal strategy $q^*(x)$

↓ View the solution

Coarse Decision Space

Assume weather classes: $K = \{\text{sunny}, \text{cloudy}, \text{rain}, \text{hailstorm}\}$. You want to go for a walk, but plan to stay inside if the weather is not $k = \text{sunny}$. Given a measurement x from your UltimateWeatherSensor, you calculated the posterior probabilities of the current weather as $p_{K|X}(\cdot \mid x) = (0.4, 0.2, 0.2, 0.2)$. The task is to decide whether it is sunny, $D \in \{\text{sunny}, \text{not sunny}\}$. What is the optimal Bayesian decision in the following cases (explain):

1. The cost of a correct decision is zero and the cost of a wrong decision is a constant $C > 0$ (normal person).
2. Mistakenly deciding $d = \text{sunny}$ costs twice less than mistakenly deciding $d = \text{not sunny}$ (an active person that does not care *that* much about getting wet).

↓ View the solution

Gaussian Three Classes

We need to classify objects into three classes $k \in \{1, 2, 3\}$. The classes are equally probable a priori. Observations x of objects in class 1 follow the distribution $\mathcal{N}(0, 1^2)$. Recall $\mathcal{N}(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2. Similarly, in classes 2 and 3, the observations are distributed as $\mathcal{N}(0, 2^2)$ and $\mathcal{N}(3, 2^2)$, respectively.

What is the optimal Bayesian decision $d \in \{1, 2, 3\}$ for the two observations $x = 1$ and $x = 0$ in the following cases:

- if the loss matrix is

$$W_a = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} .$$

- if the loss matrix is

$$W_b = \begin{bmatrix} 0 & 2 & 1 \\ 2 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} .$$

What is the probability of incorrect decision ($d \neq k$) for the first case and observation $x = 1$?

↓ View the solution

Bayes Classification Gaussians

Consider the problem of classification to two classes. The classes are denoted 1 and 2. The feature space (measurement) is one-dimensional, $x \in \mathbb{R}$. There is the following setting:

- prior probabilities are $p_1 = \dfrac{1}{1+e}$, $p_2 = \dfrac{e}{1+e}$,
- conditional probabilities are

$$p(x \mid 1) = \frac{1}{\sigma_1\sqrt{2\pi}}\, e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}, \qquad p(x \mid 2) = \frac{1}{\sigma_2\sqrt{2\pi}}\, e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}},$$

where $\sigma_1 = \sigma_2 = 1$, $\mu_1 = -2$, and $\mu_2 = 2$.

Consider the following cost matrix $W(k, d)$:

	d=1	d=2
k=1	0	e^5
k=2	e^2	0

where k is the class index as above and d is decision.

1. Define the Bayes risk $R(q)$.
2. Find the optimal strategy $q^*(x)$, which (under the settings as described) minimizes the Bayes risk $R(q)$. Write down the strategy in a clear, mathematically concise form.

↓ View the solution

Gaussian Conditionals

Recall the optimal decision strategy q minimizes the risk:

$$R(q) = \sum_{x \in X} \sum_{k \in K} p(x, k) W(k, q(x)).$$

Consider 0-1 loss function:

$$W(k, d) = \begin{cases} 1, & \text{if } k \neq d, \\ 0, & \text{if } k = d. \end{cases}$$

1. Prove:
$$q^*(x) = \arg\max_d p(d \mid x)$$

2. Let additionally $K = \{0, 1\}$. Prove $q^*(x)$ takes the form
$$\frac{p(x \mid k = 0)}{p(x \mid k = 1)} \lessgtr \theta$$

↓ View the solution

Error Correcting Codes

A digital signal transmitting system sends three bits over a noisy channel. When sending bit $k_i \in \{0, 1\}$, there is a chance that it will be flipped due to noise. To counter that, we decide for each two bits of information to send also a check bit, which however also has to be sent over the noisy channel. Assume that the true bits $k \in \{0, 1\}^3$ form an error correcting code where the last bit is always the sum of the first two bits modulo 2. Assume that the information bits have equal probabilities a-priori.

Suppose we have received bits $x = (0, 0, 1)$, the flipping errors are independent and chances of flipping were estimated as $\epsilon = (0.3, 0.4, 0.3)$ for the three bits, respectively.

1. Recognize which number is encoded by the first two bits.
2. Decide whether this packet of 3 bits has to be requested again considering that the cost of skipping an error is $100\times$ more than requesting to repeat the packet.

↓ View the solution

Non-Bayesian Decision Making

Minimax with Discrete Measurements

An aging short-sighted student at CTU wants to marry. He cannot afford to miss recognizing a girl when he meets her, therefore he sets the threshold on overlooking an opportunity as $\bar{\epsilon}_F = 0.2$. At the same time, he wants to minimize mis-classifying boys for girls. The exact setup is as follows:

- Hidden states:
  $K = \{\mathrm{F}, \mathrm{M}\}$ (female, male)
- Measurements:
  $X = \{\text{short}, \text{normal}, \text{tall}\} \times \{\text{ultralight}, \text{light}, \text{avg}, \text{heavy}\}$
- Prior probabilities are not known
- Conditional probabilities $p(x|k)$ are given in Table 1 and Table 2.

Table 1: $p(x|\mathrm{F})$

	Ultralight	Light	Avg	Heavy
Short	0.197	0.145	0.094	0.017
Normal	0.077	0.299	0.145	0.017
Tall	0.001	0.008	0.000	0.000

Table 2: $p(x|\mathrm{M})$

	Ultralight	Light	Avg	Heavy
Short	0.011	0.005	0.011	0.011
Normal	0.005	0.071	0.408	0.038
Tall	0.002	0.014	0.255	0.169

Find the optimal strategy when you formulate the task as a minimax problem

↓ View the solution

Neyman-Pearson with Discrete Measurements

An aging student at CTU wants to marry. He cannot afford to miss recognizing a girl when he meets her, therefore he sets the threshold on overlooking an opportunity as $\bar{\epsilon}_D = 0.2$. At the same time, he wants to minimize mis-classifying boys for girls. The exact setup is as follows:

- Hidden states:
  $K = \{\mathrm{D}, \mathrm{N}\} \equiv \{\mathrm{F}, \mathrm{M}\}$ (female, male)
- Measurements:
  $X = \{\text{short}, \text{normal}, \text{tall}\} \times \{\text{ultralight}, \text{light}, \text{avg}, \text{heavy}\}$
- Prior probabilities are unknown
- Conditional probabilities $p(x|k)$ are given in Table 1 and Table 2.

Table 1: $p(x|\mathrm{F})$

	Ultralight	Light	Avg	Heavy
Short	0.197	0.145	0.094	0.017
Normal	0.077	0.299	0.145	0.017
Tall	0.001	0.008	0.000	0.000

Table 2: $p(x|\mathrm{M})$

	Ultralight	Light	Avg	Heavy
Short	0.011	0.005	0.011	0.011
Normal	0.005	0.071	0.408	0.038
Tall	0.002	0.014	0.255	0.169

Find the optimal strategy when you formulate the task as a Neuman-Pearson problem.

↓ View the solution

Minimax - Continuous Measurements

Suppose that you have a two-class decision problem $y \in \{1, 2\}$ with real-valued features $x \in \langle -1, 1 \rangle$ and that only the class conditional probabilities $p(x \mid y{=}1) = \max(-x, x) = |x|$ and $p(x \mid y{=}2) = \min(1 + x, 1 - x) = 1 - |x|$ are given.

1. Write down formally the Minimax problem formulation.
2. How many thresholds does the optimal decision strategy need in the original feature space? Why?
3. Find the optimal Minimax strategy for this decision problem. Do it formally, e.g. purely graphical solution is not acceptable.

↓ View the solution

Neyman-Pearson: Continuous Measurements

Suppose that you have a two-class decision problem $y \in \{1, 2\}$ with real-valued features $x \in [0, 1]$ and that only the class conditional probabilities $p(x \mid y = 1) = 1$ and $p(x \mid y = 2) = x + 0.5$ are given.

1. Write down formally the Neyman–Pearson problem formulation.
2. Find the optimal Neyman–Pearson strategy for this decision problem when $y = 2$ is the dangerous state and the probability of overlooked danger shouldn't be higher than 0.1.

↓ View the solution

Non-Parametric Density Estimation

Parzen Window Density Estimation

Given the measurements $X = \{1.5, -1, 1.5, 3, 2, -0.5\}$, plot the non-parametric estimate of a distribution $p(x)$ using the Parzen window method with a kernel function $K(x, y) = k(x - y)$ and $k(z)$ defined as:

$$\begin{aligned} k(z) &= 1/h \qquad \text{for } |z| \leq h/2 \,, \\ k(z) &= 0 \qquad \text{for } |z| > h/2 \,, \end{aligned}$$

for $h = 3$.

What is the probability of $p(x = 2.5)$?

↓ View the solution

k-NN Classification

With the following training set with data points (x, y), where $x \in \mathbb{R}$ are the measurements and $y \in \{A, B\}$ their class, classify the point $x = 5$ using 1-NN, 3-NN and 5-NN classifier.

$$\mathcal{T} = \{(0, A), (-1.5, A), (10, B), (2, A), (4.5, A), (3, B), (6, B), (9, B), (1.5, A), (11, B)\}$$

↓ View the solution

Support Vector Machines

Hard Margin SVM

1. For a hard-margin linear SVM classifier, formally write down the primal optimisation task.
2. How are the margin size and the norm of the weight vector $||\mathbf{w}||$ related in hard-margin SVM?

↓ View the solution

Soft Margin SVM

1. For a soft-margin SVM classifier, formally write down the primal optimisation task.
2. Explain the concept of "slack variables" in the soft-margin SVM task.
3. What is the influence of the parameter C in soft-margin SVM on the decision boundary?

↓ View the solution

Graphical Interpretation

Visualize a soft-margin SVM in 2D on a toy dataset. Create a drawing that includes:

- a separating hyperplane,
- the classification strategy
- at least 6 data points,
- the margin boundaries,
- the support vectors,
- an illustration of slack variables for any points that violate the margin.

↓ View the solution

Dual Task

Consider the dual problem for a soft-margin SVM,

$$\begin{aligned} \alpha = \arg\max_{\alpha} \{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \} \\ \text{s.t.} \quad 0 \leq \alpha_i \leq C, \quad i = 1, 2, \ldots, n \\ \sum_{i=1}^{n} \alpha_i y_i = 0, \end{aligned}$$

where $\mathbf{x}_i$ are the data points, y_i are the class labels, and C is a constant.

1. What is the advantage of using the dual formulation for the SVM problem?
2. What is the meaning of the optimized variables, α_i?
3. How does the use of a kernel change the dual problem shown above, and what is the purpose of using a kernel?
4. What is the role of support vectors in the SVM dual problem?
5. How does the number of support vectors affect the complexity and generalization of the model?

↓ View the solution

Kernels

1. Explain why it is advantageous to use kernels in SVM.
2. For each function $K(x, y)$ below, determine if it is a valid SVM kernel, where $x, y \in \mathbb{R}^2$. If it is a valid kernel, find the corresponding feature mapping $\Phi(x)$. If not, explain why it is not a valid kernel.
    1. $K(x, y) = 2x - y$
    2. $K(x, y) = \sqrt{2}x - y$
    3. $K(x, y) = x^T(3y)$

↓ View the solution

Neural Networks

Basic concepts

Training

Find the most consistent matching of concepts on the left and descriptions on the right.

concept	description
1. Backpropagation	A. A way to learn neural networks
2. Gradient	B. Method to optimize training loss
3. Chain rule	C. Is necessary to find a step direction for gradient descent
4. Training loss minimization	D. A rule to compute gradient of composite functions
5. SGD	E. Computationally efficient automatic differentiation for scalar-valued composite functions

Basic math

Let $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$. Find the most consistent matching of concepts on the left and descriptions on the right.

concept	description
1. Gradient of f	A. A linear mapping approximating f locally around a point
2. Derivative of f	B. Expression of the derivative in coordinates as a matrix
3. Jacobian of f	C. Column vector of partial (or total) derivatives in case f is scalar-valued, i.e. $m = 1$

↓ View the solution

Computational graph

Example from labs.

Consider a simple linear neuron with one input

$$f(x) = wx + b$$

and a squared-error loss function

$$L(y, y^*) = (y - y^*)^2$$

where $y = f(x)$ is the neuron output and y^* is the target.

1. Draw the computational graph of $L(f(x), y^*)$ using basic operation nodes: multiplication (*), addition (+), subtraction (-), and square ($\cdot^2$).
2. For each node type, define:
   - The forward operation
   - The backward operation (gradient computation)
3. Using the following concrete values, compute the full forward pass to determine the loss $L(f(x), y^*)$.
   - Input $x = 2$
   - Target $y^* = 4$
   - Parameters $w = 3, b = 0$
4. Compute the full backward pass to find the gradients of the loss with respect to all parameters and the input
$$\frac{\partial L}{\partial w}, \frac{\partial L}{\partial b}, \frac{\partial L}{\partial x}$$

↓ View the solution

Solutions

Conditional Probabilities Solution

↑ View the exercise

Question 1 - marginal probabilities

$$P(k = \text{rain}) = 0.02 + 0.12 + 0.09 + 0.04 = 0.27$$
$$P(k = \text{no rain}) = 0.38 + 0.28 + 0.06 + 0.01 = 0.73$$

$$P(x = 1) = 0.02 + 0.38 = 0.40$$
$$P(x = 2) = 0.12 + 0.28 = 0.40$$
$$P(x = 3) = 0.09 + 0.06 = 0.15$$
$$P(x = 4) = 0.04 + 0.01 = 0.05$$

Question 2 - cloudiness

$$P(x \leq 2 \mid k = \text{rain}) = \frac{P(x \leq 2, k = \text{rain})}{P(k = \text{rain})}$$
$$P(x \leq 2, k = \text{rain}) = 0.02 + 0.12 = 0.14$$

Cancer Test Problem Solution

↑ View the exercise

Given the info you have, complete the table of joint probabilities $P(C, T)$ and the marginal probabilities $P(T)$.

Question 1 - Positive test

$P(y \mid +) = \frac{P(y,+)}{P(+)} = \frac{P(+|y)P(y)}{P(+|y)P(y)+P(+|n)P(n)} \approx 3.1\%$.

Question 2 - Negative test

$P(n \mid -) = \frac{P(n,-)}{P(-)} = \ldots \approx 99.9\%$.

Umbrella Rain Solution

↑ View the exercise

Question 1

$\frac{p(2,\text{rain})}{p(x=2)} = 0.12/(0.12+0.28) = 30\%$

Question 2

$\sum_x x \frac{p(x,\text{rain})}{p(\text{rain})} = 1 \cdot \frac{0.02}{0.27} + 2 \cdot \frac{0.12}{0.27} + 3 \cdot \frac{0.09}{0.27} + 4 \cdot \frac{0.04}{0.27} \approx 2.56$

Question 3

$R(x,d) = \sum_k W(k,d)p(k \mid x) = 10 \cdot (0.12/0.40) + (-2) \cdot (0.28/0.40) = 1.6$

Question 4

$10 \cdot p(\text{rain}) + (-2) \cdot p(\text{no rain}) = 10 \cdot 0.27 + (-2) \cdot 0.73 = 1.24$

Question 5

$0 \cdot 0.27 + 5 \cdot 0.73 = 3.65$

Question 6

Let's do it case by case. First compute partial risks $R(x,d) = \sum_k p(k \mid x)W(k,d)$ for $x = 1$.

$$\begin{aligned} R(1,\text{umbrella}) &= p(\text{rain} \mid 1)W(\text{rain},\text{umbrella}) + p(\text{no rain} \mid 1)W(\text{no rain},\text{umbrella}) \\ &= 0.05 \cdot 0 + 0.95 \cdot 5 = 4.75 \end{aligned}$$

$$\begin{aligned} R(1,\text{no umbrella}) &= p(\text{rain} \mid 1)W(\text{rain},\text{no umbrella}) + p(\text{no rain} \mid 1)W(\text{no rain},\text{no umbrella}) \\ &= 0.05 \cdot 10 + 0.95 \cdot -2 = -1.4 \end{aligned}$$

$$\begin{aligned} R(1,100) &= p(\text{rain} \mid 1)W(\text{rain},100) + p(\text{no rain} \mid 1)W(\text{no rain},100) \\ &= 0.05 \cdot 5 + 0.95 \cdot 0 = 0.25 \end{aligned}$$

Since $R(1,\text{no umbrella}) < R(1,100) < R(1,\text{umbrella})$ we decide $q^*(1) = \text{no umbrella}$.

Similarly for the other measurements.

- $q^*(2) = 100$ (partial risk 1.5)
- $q^*(3) = \text{umbrella}$ (partial risk 2)
- $q^*(4) = \text{umbrella}$ (partial risk 1).

Coarse Decision Space Solution

↑ View the exercise

Question 1

The optimal decision is *not sunny* (because $0.4 \cdot C < (0.2 + 0.2 + 0.2) \cdot C$).

Question 2

The optimal decision is *sunny* (because $(0.2 + 0.2 + 0.2) \cdot \frac{C}{2} < 0.4 \cdot C$).

Gaussian Three Classes Solution

↑ View the exercise

Lets first compute the probability densities for the given measurements $x \in \{0, 1\}$ by evaluating the normal distributions $p(x \mid k)$ (approximately).

	$k = 1$	$k = 2$	$k = 3$
$x = 0$	0.399	0.199	0.065
$x = 1$	0.242	0.176	0.121

Then we convert to the joint $p(x, k)$ (multiply by $p(k)$).

	$k = 1$	$k = 2$	$k = 3$
$x = 0$	0.133	0.066	0.022
$x = 1$	0.081	0.059	0.040

Finally we divide by $p(x) = p(x, 1) + p(x, 2) + p(x, 3)$ to get the posterior $p(k \mid x)$.

	$k = 1$	$k = 2$	$k = 3$
$x = 0$	0.602	0.299	0.099
$x = 1$	0.450	0.328	0.222

In the first part with loss W_a (0–1 loss), we can simply pick the maximum posterior, resulting in $d = 1$ for both $x = 0$ and $x = 1$.

In the second part with loss W_b we compute the partial risks $\sum_k p(k \mid x)\, W(k, d)$.

	$d = 1$	$d = 2$	$d = 3$
$x = 0$	0.697	1.303	0.901
$x = 1$	0.878	1.122	0.778

We pick the decision with minimal partial risk: $q^*(x = 0) = 1$ and $q^*(x = 1) = 3$.

Finally, we compute the probability of incorrect decision with W_a and $x = 1$. Since $q^*(x) = 1$ in this case (see above), the probability of incorrect decision is $p(k = 2 \mid x = 1) + p(k = 3 \mid x = 1) = 0.328 + 0.222 = 0.550$

Bayes Classification Gaussians Solution

↑ View the exercise

Question 1: Define the Bayes risk $R(q)$

$$R(q) = \int_{\mathbb{R}} \sum_k p(x,k)\, W(k, q(x))\, dx,$$

where $p(x,k) = p(x \mid k)\, p_k$

Question 2: Find the optimal strategy $q^*(x)$.

First we compute the two partial risks. For decision 1

$$R(x,1) = W(1,1)p(1 \mid x) + W(2,1)p(2 \mid x) = 0 \cdot p(1 \mid x) + e^2 p(2 \mid x) = e^2 p(2 \mid x),$$

and for decision 2

$$R(x,2) = W(1,2)p(1 \mid x) + W(2,2)p(2 \mid x) = e^5 p(1 \mid x) + 0 \cdot p(2 \mid x) = e^5 p(1 \mid x).$$

The optimal strategy chooses class 1 iff $R(x,1) \leq R(x,2)$:

$$e^2 p(2 \mid x) \leq e^5 p(1 \mid x) \quad \Longleftrightarrow \quad \frac{p(2 \mid x)}{p(1 \mid x)} \;\leq\; e^3.$$

Using Bayes rule,

$$\frac{p(2 \mid x)}{p(1 \mid x)} = \frac{p_2\, p(x \mid 2)}{p_1\, p(x \mid 1)}.$$

we have

$$\frac{p(x \mid 2)}{p(x \mid 1)} \leq \frac{p_1}{p_2} e^3 = \frac{1}{e} e^3 = e^2$$

Plugging in the normal distribution formulas and simplifying we get

$$
\begin{aligned}
e^{-\frac{1}{2}\left[(x-\mu_2)^2-(x-\mu_1)^2\right]} &\leq e^2 \\
-\frac{1}{2}\left[(x-2)^2-(x-(-2))^2\right] &\leq 2 \\
-\frac{1}{2}\left[x^2-4x+4-(x^2+4x+4)\right] &\leq 2 \\
-\frac{1}{2}\left[-8x\right] &\leq 2 \\
4x &\leq 2 \\
x &\leq \frac{1}{2}
\end{aligned}
$$

Final strategy

$$
\boxed{q^\star(x) = \begin{cases} 1, & x \leq \frac{1}{2}, \\ 2, & x > \frac{1}{2}. \end{cases}}
$$

Gaussian Conditionals Solution

↑ View the exercise

Throughout, assume $p(x) > 0$ for any observation x we consider (otherwise the choice of $q(x)$ on a zero-probability set is immaterial). We will make conditioning on $X = x$ explicit and derive the decision minimizing the **partial risk** at that x.

Question 1 - Under 0–1 loss, $q(x) = \arg\max_d p(d \mid x)$

We start by rewriting Bayesian risk via partial risk.

$$R(q) = \sum_x \sum_k p(x)\, p(k \mid x)\, W\Big(k, q(x)\Big) = \sum_x p(x) \underbrace{\Big(\sum_k p(k \mid x)\, W(k, q(x))\Big)}_{=:R(x,q(x))}.$$

Here $R(x, q(x))$ is the **partial risk** at x. Since $p(x) \geq 0$, minimizing $R(q)$ over q is equivalent to minimizing each $R(x, q(x))$ separately (pointwise in x).

Next we plug in the 0–1 loss. For a fixed x and decision d, the partial risk is:

$$R(x, d) = \sum_k p(k \mid x)\, W(k, d) = p(d \mid x) \cdot 0 + \sum_{k \neq d} p(k \mid x) \cdot 1 = \sum_k p(k \mid x) - p(d \mid x).$$

Because $\sum_k p(k \mid x) = 1$, this simplifies to

$$R(x, d) = 1 - p(d \mid x).$$

Finally, minimizing $R(x, d) = 1 - p(d \mid x)$ over d is equivalent to maximizing $p(d \mid x)$ over d. Therefore

$$q^*(x) = \arg\min_d R(x, d) = \arg\max_d p(d \mid x).$$

Question 2 - Binary case $K = \{0, 1\}$: likelihood-ratio form

From the solution to question 1 we know that we decide 0 iff $p(0 \mid x) \geq p(1 \mid x)$. Using Bayes' rule,

$$p(0 \mid x) = \frac{p(x \mid 0)p(0)}{p(x)}, \qquad p(1 \mid x) = \frac{p(x \mid 1)p(1)}{p(x)}.$$

Thus,

$$p(0 \mid x) \geq p(1 \mid x) \iff p(x \mid 0)p(0) \geq p(x \mid 1)p(1) \iff \frac{p(x \mid 0)}{p(x \mid 1)} \geq \frac{p(1)}{p(0)}.$$

Therefore the optimal rule has the likelihood-ratio form

$$\frac{p(x \mid k = 0)}{p(x \mid k = 1)} \gtrless \theta, \qquad \text{with } \theta = \frac{p(1)}{p(0)}.$$

Error Correcting Codes Solution

↑ View the exercise

Question 1

We observe a sequence x of transmitted over a noisy channel sequence of the true bits k. We have

$$\begin{aligned} p(x_1{=}0 \mid k_1{=}0) &= 1 - \epsilon_1 = 0.7 \\ p(x_2{=}0 \mid k_2{=}0) &= 1 - \epsilon_2 = 0.6 \\ p(x_3{=}1 \mid k_3{=}1) &= 1 - \epsilon_3 = 0.7. \end{aligned}$$

If we were to recognize each bit independently (i.e. via $p(k_i \mid x_i) \propto p(x_i \mid k_i)p(k_i)$, assuming $p(k_i) = 0.5$), we would recover $k = (0, 0, 1)$, which is incorrect as it could not arise as an error-correcting code.

The only possible sequences that could have been sent (with the error-correcting code) are:

$$\begin{aligned} &(0,0,0) \\ &(0,1,1) \\ &(1,0,1) \\ &(1,1,0). \end{aligned}$$

So there are only 4 possible hidden states. We need to compute and compare their posterior probabilities $p(k \mid x)$. For the purpose of selecting the best posterior, it is sufficient to compare only the numbers $p(x \mid k)$ as the hidden states are equiprobable a priori. Given the noisy channel model and the flipping probabilities we compute:

k	$p(x \mid k)$
$(0,0,0)$	$0.7 \cdot 0.6 \cdot 0.3 = 0.126$
$(0,1,1)$	$0.7 \cdot 0.4 \cdot 0.7 = 0.196$
$(1,0,1)$	$0.3 \cdot 0.6 \cdot 0.7 = 0.126$
$(1,1,0)$	$0.3 \cdot 0.4 \cdot 0.3 = 0.036$

The most likely hidden state according to this distribution is $(0, 1, 1)$.

Question 2

For this part we will need the complete probability $p(k \mid x)$. It is obtained by renormalizing the four values, i.e., dividing them by their sum, 0.484. The probability of correct decoding is thus $0.196/0.484 \approx 0.405$ and of incorrect decoding respectively $1 - 0.196/0.484 \approx 0.595$. The decision to keep the message has the (partial) risk of $0.595 \cdot 100 \cdot C$ while the decision to request a repeat has the risk of $0.405 \cdot C$ only. We decide to ask for a repeat.

Minimax - Discrete Measurements

↑ View the exercise

Let $D = \{\mathrm{M}, \mathrm{F}\}$ be the decision set and $q : X \to D$ a decision strategy. The strategy makes two types of errors:

$$\epsilon_F(q) = \sum_{x:\, q(x)\neq \mathrm{F}} p(x \mid \mathrm{F}), \qquad \epsilon_M(q) = \sum_{x:\, q(x)\neq \mathrm{M}} p(x \mid \mathrm{M}).$$

We will use the likelihood ratio

$$r(x) = \frac{p(x \mid \mathrm{M})}{p(x \mid \mathrm{F})},$$

because it is known that the optimal strategy can be expressed as

$$q^*(x) = \begin{cases} M & \text{if} \quad r(x) > \mu \\ F & \text{if} \quad r(x) < \mu \end{cases}.$$

Pre-computed quantities

We start by pre-computing several useful quantities.

Likelihood ratio $r(x) = p_M/p_F$:

Height \ Weight	Ultralight	Light	Avg	Heavy
Short	0.0559	**0.0345**	0.1171	0.6471
Normal	0.0649	0.2375	2.8138	2.2353
Tall	2.0000	1.7500	∞	∞

(∞ where $p_F = 0 < p_M$)

Rank of the measurements by $r(x)$ (1 = smallest r, i.e. strongest evidence for F):

Height \ Weight	Ultralight	Light	Avg	Heavy
Short	2	**1**	4	6
Normal	3	5	10	9
Tall	8	7	11	12

(ties among ∞ set as 11,12)

Cumulative sums when sweeping **increasing** $r(x)$. Each cell shows the cumulative sum **up to and including** that cell in the rank order:

Cumulative $\sum p(x \mid \mathbf{F})$

Height \ Weight	Ultralight	Light	Avg	Heavy
Short	0.342	**0.145**	0.513	0.829
Normal	0.419	0.812	1.000	0.855
Tall	0.838	0.837	1.000	1.000

Cumulative $\sum p(x \mid \mathbf{M})$

Height \ Weight	Ultralight	Light	Avg	Heavy
Short	0.016	**0.005**	0.032	0.114
Normal	0.021	0.103	0.576	0.168
Tall	0.130	0.128	0.831	1.000

Note: In a test it is probably better to compute these on the fly, without constructing the whole table first.

Minimax

The problem:

$$q^* = \arg \min_{q:X \to \{\mathrm{F},\mathrm{M}\}} \max\{\epsilon_D(q),\, \epsilon_N(q)\}$$

We sweep the cells in the above tables in the order of increasing $r(x)$ (strongest evidence for F first). At the beginning the strategy is to classify all the cells as M, with each step we mark one more cell to be classified as F. For example in the third iteration we will be considering the following stategy:

Height \ Weight	Ultralight	Light	Avg	Heavy
Short	**F**	**F**	M	M
Normal	**F**	M	M	M
Tall	M	M	M	M

After the first eight cells we have

Height \ Weight	Ultralight	Light	Avg	Heavy
Short	**F**	**F**	**F**	**F**
Normal	**F**	**F**	M	M
Tall	**F**	**F**	M	M

From the ranked/cumulative tables: $\sum p(x|F) = 0.838$, $\sum p(x|M) = 0.130$ thus $\epsilon_F = 1 - 0.838 = 0.162$, and $\epsilon_M = 0.130$. The maximum of these two errors is 0.162. In each of these eight steps the maximum of the errors decreases (we are finding better solutions).

If we also include cell nine:

Height \ Weight	Ultralight	Light	Avg	Heavy
Short	**F**	**F**	**F**	**F**
Normal	**F**	**F**	M	**F**
Tall	**F**	**F**	M	M

we get $\sum p(x|F) = 0.855$ and $\sum p(x|M) = 0.168$, thus $\epsilon_F = 0.145$, $\epsilon_M = 0.168$, and their maximum is 0.168. Here, the maximum *increased* for the first time (the solution becomes worse than the previous one). Continuing the procedure would increase ϵ_M in each step, so there is no hope to find the optimal strategy in the following iterations.

Thus the **optimal minimax strategy** predicts F for the first eight cells and M for the rest of the cells. The achieved minimax error is 0.162. The strategy is shown in the following table:

Height \ Weight	Ultralight	Light	Avg	Heavy
Short	**F**	**F**	**F**	**F**
Normal	**F**	**F**	M	M
Tall	**F**	**F**	M	M

Note: If we allow the strategy to be randomized (randomly (in $\approx 58\%$ of cases) selecting F for the cell with rank 9), we may get even lower minmax error ≈ 0.1521.

Neyman-Pearson - Discrete Measurements

↑ View the exercise

Let $D = \{\mathrm{M}, \mathrm{F}\}$ be the decision set and $q : X \to D$ the decision strategy. The strategy makes two types of errors:

$$\epsilon_F(q) = \sum_{x:\, q(x)\neq \mathrm{F}} p(x \mid \mathrm{F}), \qquad \epsilon_M(q) = \sum_{x:\, q(x)\neq \mathrm{M}} p(x \mid \mathrm{M}).$$

We will use the likelihood ratio

$$r(x) = \frac{p(x \mid \mathrm{M})}{p(x \mid \mathrm{F})} \, .$$

because it is known that the optimal strategy can be expressed as

$$q^*(x) = \begin{cases} M & \text{if} \quad r(x) > \mu \\ F & \text{if} \quad r(x) < \mu \end{cases} .$$

Pre-computed quantities

We start by pre-computing several quantities.

Likelihood ratio $r(x) = p_M/p_F$:

Height \ Weight	Ultralight	Light	Avg	Heavy
Short	0.0559	**0.0345**	0.1171	0.6471
Normal	0.0649	0.2375	2.8138	2.2353
Tall	2.0000	1.7500	∞	∞

(∞ where $p_F = 0 < p_M$)

Rank of the measurements by $r(x)$ (1 = smallest r, i.e. strongest evidence for F):

Height \ Weight	Ultralight	Light	Avg	Heavy
Short	2	**1**	4	6
Normal	3	5	10	9
Tall	8	7	11	12

(ties among ∞ set as 11,12)

Cumulative sums when sweeping **increasing** $r(x)$. Each cell shows the cumulative sum **up to and including** that cell in the rank order:

Cumulative $\sum p(x \mid \mathbf{F})$

Height \ Weight	Ultralight	Light	Avg	Heavy
Short	0.342	**0.145**	0.513	0.829
Normal	0.419	0.812	1.000	0.855
Tall	0.838	0.837	1.000	1.000

Cumulative $\sum p(x \mid \mathbf{M})$

Height \ Weight	Ultralight	Light	Avg	Heavy
Short	0.016	**0.005**	0.032	0.114
Normal	0.021	0.103	0.576	0.168
Tall	0.130	0.128	0.831	1.000

Note: In a test it is probably better to compute these on the fly, without constructing the whole table first.

Neyman-Pearson

The problem:

$$\min_{q:X\to D} \epsilon_M(q)$$
$$\text{s.t.} \quad \epsilon_F(q) \le \bar{\epsilon}_D = 0.2 \quad \Big(\iff \sum_{x:q(x)=\mathrm{F}} p(x \mid \mathrm{F}) > 0.8 \Big).$$

For simplicity, we consider only deterministic strategies, i.e. given the measurement the strategy decides deterministically to either of the classes. The optimal strategy is found by taking the **lowest** $r(x)$ cells and marking them as F until the constraint is satisfied. (This corresponds to systematically evaluating all likelihood ratio thresholds μ).

From the cumulative tables, the first 5 cells (by rank) give $\sum p_F = 0.812 > 0.8$.

So the optimal Neyman-Pearson (deterministic) strategy is

Height \ Weight	Ultralight	Light	Avg	Heavy
Short	**F**	**F**	**F**	M
Normal	**F**	**F**	M	M
Tall	M	M	M	M

Errors achieved (deterministic NP)

- $\epsilon_F = 1 - 0.812 = \mathbf{0.188} \ \le 0.2$ (constraint satisfied).
- $\epsilon_M = 0.005 + 0.011 + 0.005 + 0.011 + 0.071 = \mathbf{0.103}$.

Note: allowing randomization would tighten ϵ_F exactly to 0.2 and slightly reduce ϵ_M to ≈ 0.1002.

Minimax - Continuous Measurements

↑ View the exercise

Question 1 - minimax definition

We define the objective function of the minimax task as

$$q^* = \arg\min_{q:\, X \to Y} \max_{y \in Y} \sum_{y \in Y} \epsilon(y) \quad \text{, where} \epsilon(y) = \sum_{x:\ q(x) \neq y} p(x \mid y),$$

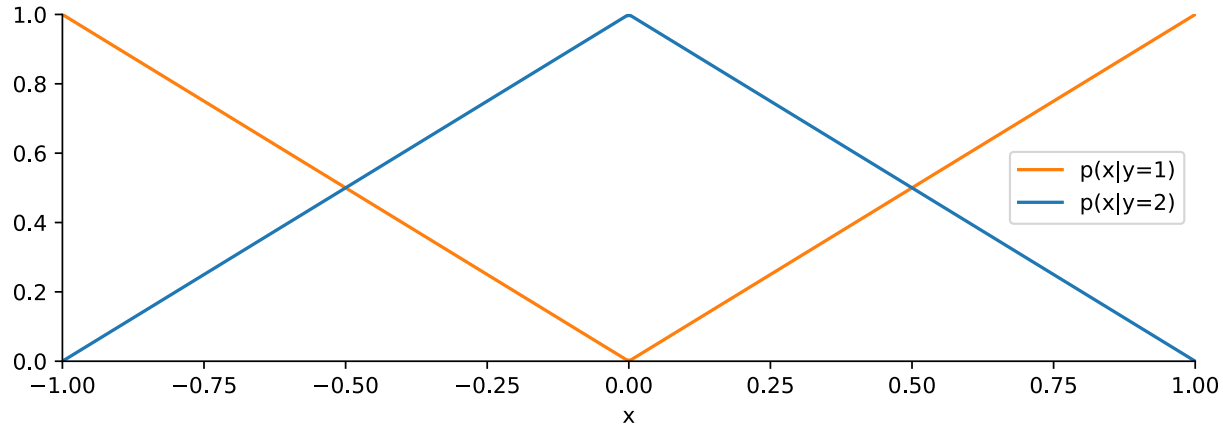
where $Y = \{1, 2, \ldots, N\}$ are classes, X is the set of observations x, $p(x \mid y)$ are conditionals that are known $\forall y \in Y$, and $q : X \to Y$ is a decision strategy.

Question 2 - how many thresholds on x?

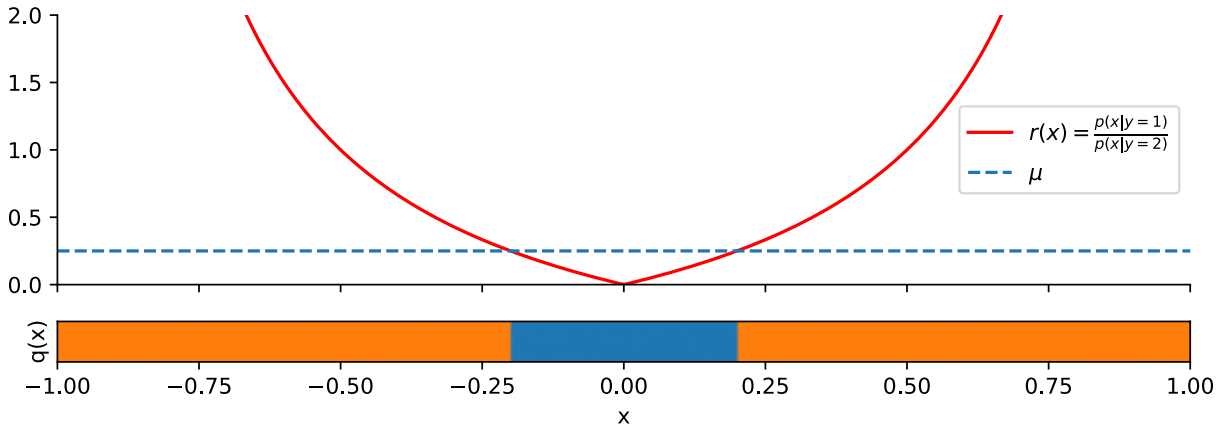
For a 2-class Minimax problem the solution is always defined by a single likelihood ratio threshold μ.

$$q^*(x) = \begin{cases} 1 & \text{if} \quad r(x) > \mu \\ 2 & \text{if} \quad r(x) < \mu \end{cases}.$$

Lets plot the probability distributions



and the likelihood ratio $r(x) = \frac{p(x|y=1)}{p(x|y=2)}$



From the plot we see that the threshold on the likelihood ratio induces a strategy that can be represented by two thresholds in the original space. Moreover, the solution will be symmetric in this particular instance, which is something we may use to simplify the derivation of the solution.

Question 3 - optimal strategy

The minimax task

$$q^* = \arg \min_{q:\, X \to Y} \max_{y \in Y} \sum_{x:\ q(x) \neq y} p(x \mid y),$$

can be rewritten as

$$q^* = \arg \min_{q(x)} \max \left\{ \int_{X_2} p(x \mid y = 1)\, dx,\ \int_{X_1} p(x \mid y = 2)\, dx \right\}$$

where

$$q(x) = \begin{cases} 1, & x \in X_1 \\ 2, & x \in X_2 \end{cases},$$
$$X_1 \cup X_2 = X,$$
$$X_1 \cap X_2 = \emptyset.$$

From the Question 2 we know that we are looking for $t \in \langle -1, 1 \rangle$ such that

$$X_1 = \langle -1, -t \rangle \cup \langle t, 1 \rangle,$$
$$X_2 = \langle -t, t \rangle.$$

Therefore the task becomes

$$q^* = \arg \min_{q(x)} \max \left\{ \int_{-t}^{t} p(x \mid y = 1)\, dx,\ \int_{-1}^{-t} p(x \mid y = 2)\, dx + \int_{t}^{1} p(x \mid y = 2)\, dx \right\}.$$

Because the likelihood ratio $r(x)$ is symmetrical around $x = 0$ we can simplify that to

$$q^* = \arg \min_{q(x)} \max \left\{ \int_{-t}^{0} p(x \mid y = 1)\, dx,\ \int_{-1}^{-t} p(x \mid y = 2) \right\}.$$

So, finding the optimal q^* corresponds to the following condition:

$$\int_{-t}^{0} p(x \mid y=1)\,dx = \int_{-1}^{-t} p(x \mid y=2)\,dx$$

(if we changed the t a little bit to either side, one of the errors would increase, resulting in worse minmax score)

Let's solve this equation:

$$\begin{aligned}
\int_{-t}^{0} -x\,dx &= \int_{-1}^{-t} 1+x\,dx \\
-\left[\frac{x^2}{2}\right]_{-t}^{0} &= \left[x+\frac{x^2}{2}\right]_{-1}^{-t} \\
\frac{t^2}{2} &= \frac{t^2}{2} - t + 1 - \frac{1}{2} \\
t &= \frac{1}{2}
\end{aligned}$$

The optimal strategy is

$$q^*(x) = \begin{cases} 1, & \text{if} \quad x \in \langle -1, -\frac{1}{2}\rangle \cup \langle \frac{1}{2}, 1\rangle \\ 2, & \text{otherwise} \end{cases}$$

Neyman-Pearson - Continuous Measurements

↑ View the exercise

Question 1 - problem formulation

Notice, that the dangerous and safe states are swapped compared to the lecture slides.

Let a decision strategy $q(x) \in \{1, 2\}$ output class 1 (safe) or class 2 (danger).

Define the two conditional error probabilities:

- **False alarm**:
$$\epsilon_1(q) = \int_{x:\ q(x)\neq 1} p(x \mid 1)$$
- **Overlooked danger**:
$$\epsilon_2(q) = \int_{x:\ q(x)\neq 2} p(x \mid 2)$$

The Neyman-Pearson problem is then defined as

$$\begin{aligned} q^* &= \arg\min_q \epsilon_1(q) \\ \text{s. t. } \epsilon_2(q) &\leq \bar{\epsilon}_2 \quad (\text{with } \bar{\epsilon}_2 = 0.1). \end{aligned}$$

Question 2 - optimal strategy

Compute the likelihood ratio and check its (non-)monotonicity

$$r(x) = \frac{p(x \mid 2)}{p(x \mid 1)} = \frac{x + 0.5}{1} = x + 0.5, \quad x \in [0, 1].$$

$r(x)$ is strictly increasing in x. Therefore $\{x : r(x) \geq \eta\}$ is an interval of the form $[c, 1]$ with $c = \eta - 0.5$.

Enforce the missed-danger constraint $\epsilon_2 \leq 0.1$ The decision rule has the form:

$$q(x; c) = \begin{cases} 2 & \text{if } x \geq c, \\ 1 & \text{if } x < c, \end{cases} \quad \text{for some } c \in [0, 1].$$

The missed danger probability is

$$\epsilon_2(q(\cdot; c)) = \mathbb{P}(q(X; c) = 1 \mid y = 2) = \int_0^c (x + 0.5)\, dx = \left[\frac{x^2}{2} + 0.5x\right]_0^c = \frac{c^2}{2} + 0.5c.$$

Set $\epsilon_2(q(\cdot; c)) = 0.1$ and solve for c:

$$\frac{c^2}{2} + 0.5c = 0.1 \iff c^2 + c = 0.2 = \frac{1}{5} \iff c^2 + c - \frac{1}{5} = 0.$$

Quadratic formula (the other solution is outside the $[0, 1]$ feature domain and makes no sense):

$$c = \frac{-1 + \sqrt{1 + \frac{4}{5}}}{2} = \frac{-1 + \sqrt{\frac{9}{5}}}{2} = \frac{-1 + \frac{3}{\sqrt{5}}}{2} \approx 0.17082.$$

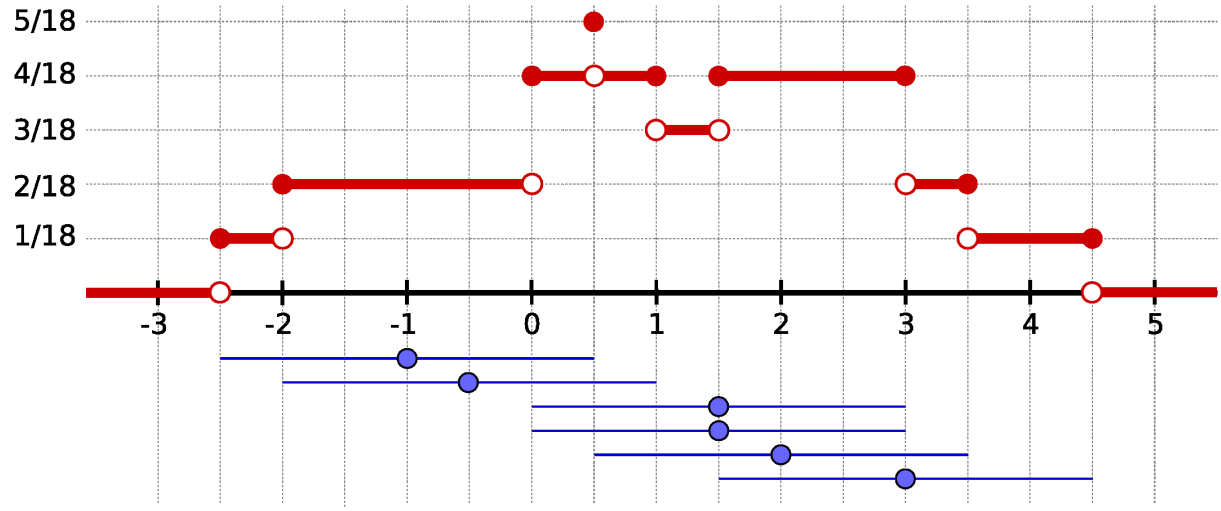
Thus the Neyman-Pearson-optimal strategy (with $\bar{\epsilon}_2 = 0.1$) is:

$$\boxed{q^*(x) = \begin{cases} 2 & \text{if } x \geq 0.17082\ldots, \\ 1 & \text{otherwise.} \end{cases}}$$

Parzen Window Density Estimation Solution

↑ View the exercise

The figure below shows the kernel density estimate in red. At the bottom (in blue) are shown copies of the kernels placed at data points. Note that some isolated points jump up because the kernel is uniform over a *closed* interval. However, these points are not important for the resulting distribution because they have a zero measure.



Construction:

1. Mark the data points $x_i \in X$ at their coordinates and center a kernel $k(x_i - x)$ at these points. In the above figure, the data points are visualized as blue circles with black outlines (shifted below the x axis for clarity) and kernels as blue lines centered at data points. The size of the kernel is $\frac{h}{2}$ to each side of the data point. The dataset contains $N = 6$ points.

2. Next we sum up the number of overlapping kernels for each x (e.g. at point -1.5 there are two overlapping kernels).

3. To compute the true value of the density estimate $\hat{p}(x)$ at any point x we use the formula for Parzen estimate:

$$\hat{p}(x) = \frac{1}{N} \sum_{i=1}^{N} K(x, x_i) \ ,$$

   where $K(x, x_i)$ is either $1/h$ or zero. The number of overlapping kernels computed in the previous step tells us how many non-zero elements are in this sum. So, for instance, if there are two overlapping kernels (as for $x = -1.5$ above), the value of $\hat{p}(x)$ is $2/(6 \cdot 3) = 2/18$

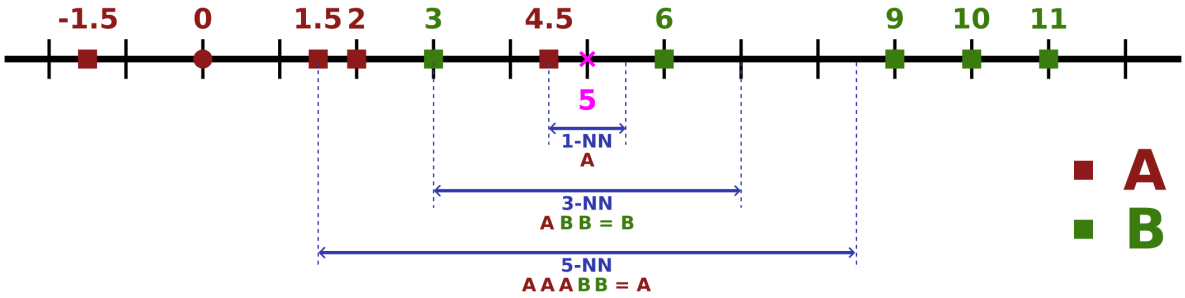
4. Do this for every $x \in \mathbb{R}$ and plot the $\hat{p}(x)$. Notice that there may be point-wise discontinuities at the points where "edges" of kernels meet.

The probability at point $x = 2.5$ is thus

$$p(x = 2.5) = \frac{4}{18} .$$

k-NN Classification Solution

↑ View the exercise



See the figure above. Notes:

- 1-NN: we are looking for the closest data point (4.5, A) which gives us the classification
- 3-NN: we consider three closest data points – the third closest is (3, B) – and the majority vote decides the classification (the same principle for the 5-NN)

Hard Margin SVM Solution

↑ View the exercise

Question 1 - Primal optimisation task

For training data $T = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, containing D-dimensional observations $x_i \in \mathbb{R}^D$ and the corresponding classes $y_i \in \{1, -1\}$, we search for

$$\begin{aligned} \mathbf{w}^*, b^* = \arg\min_{\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{s.t.} \quad & y_i\Big(\mathbf{w}\cdot\mathbf{x}_i + b\Big) \geq 1, \qquad i = 1, \ldots, n. \end{aligned}$$

- Minimizing the objective $\frac{1}{2}\|\mathbf{w}\|^2$ widens the margin.
- The constraints enforce correct classification with margin.

Question 2 - Relationship of margin and $\|\mathbf{w}\|$

The two margin hyperplanes are

$$\mathbf{w}\cdot\mathbf{x} + b = +1 \quad \text{and} \quad \mathbf{w}\cdot\mathbf{x} + b = -1,$$

and their perpendicular distance (i.e., **margin width**) is

$$m = \frac{2}{\|\mathbf{w}\|}.$$

Hence, larger margin $\Leftrightarrow$ smaller $\|\mathbf{w}\|$. The optimisation indeed maximalizes the margin by minimising $\|\mathbf{w}\|^2$ (and thus also minimizing $\|\mathbf{w}\|$).

Soft Margin SVM Solution

↑ View the exercise

Question 1 - Primal optimisation task

For training data $T = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, containing D-dimensional observations $x_i \in \mathbb{R}^D$ and the corresponding classes $y_i \in \{1, -1\}$, we search for

$$\begin{aligned} &\min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{\|\mathbf{w}\|^2}{2} + C\sum_{i=1}^{n} \xi_i \\ \text{s.t.} \quad & y_i\Big(\mathbf{w}\cdot\mathbf{x}_i + b\Big) \geq 1 - \xi_i, \qquad i = 1, \ldots, n, \\ & \xi_i \geq 0. \end{aligned}$$

The first term of the optimized function controls margin size (via $\|\mathbf{w}\|$). The second term penalises margin violations through the ξ_i, with trade-off hyper-parameter $C > 0$.

Question 2 - What are slack variables?

Slack variables $\xi_i \geq 0$ are additional optimization variables (one per constraint equation) allow for margin violation ($1 - \xi_i$ instead of 1 in the hard margin constraint).

After optimization, each ξ_i measures how much the i-th example violates the unit-margin constraint. Interpretation: * $\xi_i = 0$: correctly classified and ouside of the margin strip * $0 < \xi_i < 1$: correctly classified but *inside* the margin. * $\xi_i \geq 1$: misclassified point (lies on the wrong side of the decision boundary).

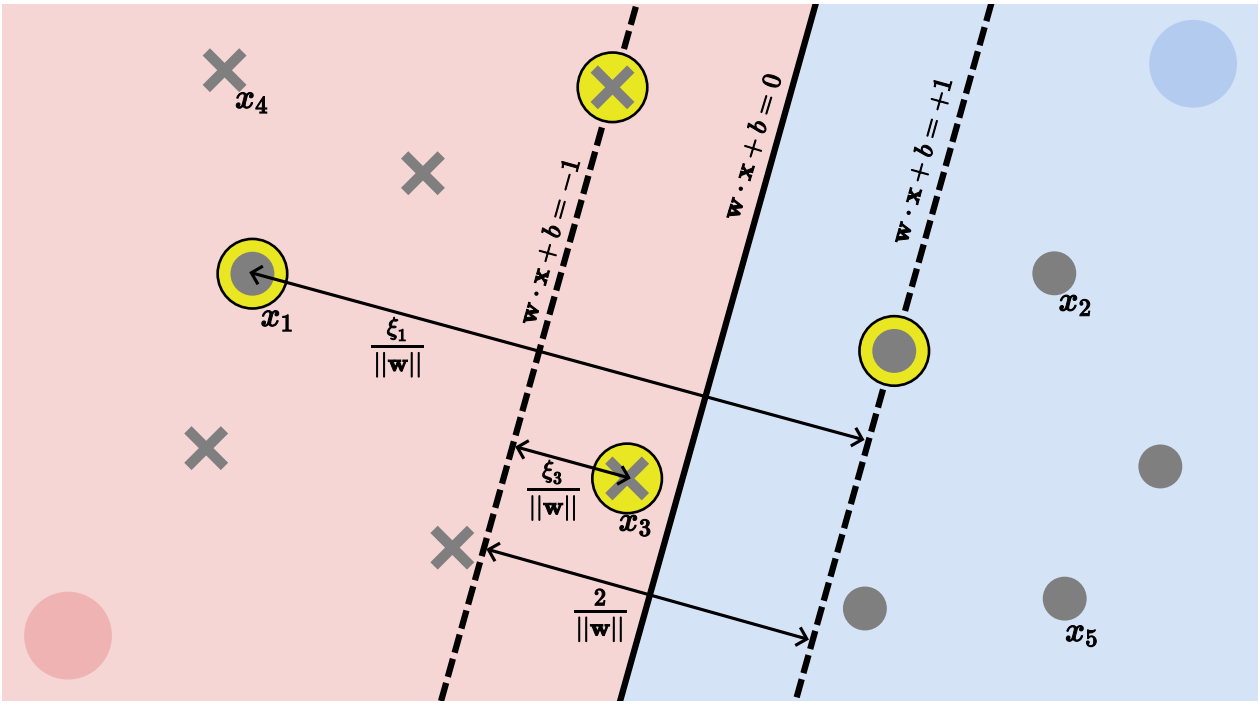
Question 3 - Influence of C on the decision boundary

C is the penalty paid by the optimization for violations of the constraints. It controls the trade-off between margin maximization and constraints violations:

- **Large** $C \Rightarrow$ violations are expensive $\Rightarrow$ the optimiser prefers smaller ξ_i, even if that requires larger $\|\mathbf{w}\|$ (narrower margin). Tends toward fitting training data more tightly (risk of overfitting).
- **Small** $C \Rightarrow$ violations are cheaper $\Rightarrow$ the optimiser allows larger ξ_i in exchange for smaller $\|\mathbf{w}\|$ (wider margin). This yields a **smoother** boundary (risk of **underfitting** if too small).

Graphical Interpretation Solution

↑ View the exercise



Support vectors are highlighted in yellow.

Dual Task Solution

↑ View the exercise

Question 1 - Advantages of the dual formulation

Key advantages:

1. **Kernelization (inner products only).** The dual depends on the data only through inner products $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$. This allows the **kernel trick**: replace $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$ by $K(\mathbf{x}_i, \mathbf{x}_j)$, enabling **non-linear** decision boundaries in the original input space without ever computing coordinates in a high-dimensional feature space.
2. **Sparsity of the solution.** In the optimum, most $\alpha_i = 0$. Only the **support vectors** ($\alpha_i > 0$) determine $\mathbf{w}$ and the classifier. This yields compact models and fast prediction when the number of support vectors is small.

Question 2 - Meaning of the variables α_i

Each α_i is a **Lagrange multiplier** attached to the margin constraint

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i.$$

They weight the contribution of training examples to the normal vector:

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i.$$

Their values encode the geometric role of each point (via KKT complementarity):

α_i value	Geometric status of $\mathbf{x}_i$	Typical condition
$\alpha_i = 0$	Not a support vector; safely outside margin	$y_i f(\mathbf{x}_i) > 1$
$0 < \alpha_i < C$	On the margin (support vector)	$y_i f(\mathbf{x}_i) = 1,\ \xi_i = 0$
$\alpha_i = C$	Inside the margin or misclassified (support vector)	$y_i f(\mathbf{x}_i) < 1,\ \xi_i > 0$

Thus, the α_i directly indicate which points are influential and *how*.

Question 3 - Effect and purpose of kernels

Effect in the dual: replace inner products by a kernel:

$$\langle \mathbf{x}_i, \mathbf{x}_j \rangle \longrightarrow K(\mathbf{x}_i, \mathbf{x}_j),$$

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)$$

with the same constraints $0 \leq \alpha_i \leq C$ and $\sum_i \alpha_i y_i = 0$.

Purpose: kernels realize an implicit feature map $\phi(\mathbf{x})$ with $K(\mathbf{x}, \mathbf{z}) = \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle$. This yields **non-linear classifiers** in the original input space while keeping optimization entirely in terms of kernel evaluations. The decision function becomes

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i y_i \, K(\mathbf{x}_i, \mathbf{x}) + b.$$

Question 4 - Role of support vectors

Only training points with $\alpha_i > 0$ (the support vectors) appear in $\mathbf{w}$ and in $f(\mathbf{x})$. They define the separating hyperplane and the margin

$$\mathbf{w} \cdot \mathbf{x} + b = 0, \qquad \mathbf{w} \cdot \mathbf{x} + b = \pm 1.$$

Once we solve the dual and get α_i, the primal weight vector is

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i.$$

Only the support vectors (samples with $\alpha_i > 0$) contribute to $\mathbf{w}$.

A convenient way to compute b is to use any support vector $\mathbf{x}_S$ with $0 < \alpha_S < C$ - a margin SV (if there is one):

$$b = y_S - \sum_{j=1}^{n} \alpha_j y_j \, K(\mathbf{x}_j, \mathbf{x}_S).$$

Note: A more complex solution for b is needed if there is no such support vector. See the explanation in the lab assignment.

Question 5 - Number of support vectors: complexity & generalization

- **Prediction cost** is $O(\#\text{SV})$ per test point, since $f(\mathbf{x}) = \sum_{i \in \mathcal{S}} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b$ sums only over the support vectors $\mathcal{S}$. So, more SVs $\Rightarrow$ more complex model (when using kernels), slower prediction and larger memory.
- **Generalization.** Fewer SVs typically indicate a larger margin and cleaner separation, often correlating with better generalization. More SVs often signal a more complex decision boundary and potential overfitting.

Kernels Solution

↑ View the exercise

Question 1 - Why kernels are advantageous

The SVM dual task formulation depends on data only via dot products $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$. Replacing them by a kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle$ yields a classifier that is linear in a (possibly very high-dimensional) feature space but non-linear in the original input space, without ever computing Φ explicitly (the **kernel trick**).

This enables rich non-linear decision boundaries with the same convex optimisation machinery.

Question 2 - Validity and feature maps

A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a valid kernel iff $K(x, y) = \langle \Phi(x), \Phi(y) \rangle$ for some (possibly infinite-dimensional) feature map Φ. It follows that a kernel must return a **scalar** and be **symmetric** in x, y (not the only conditions!); vector-valued outputs or non-symmetric maps are **not** kernels.

1) $K(x, y) = 2x - y$

- Here $x, y \in \mathbb{R}^2$. The expression $2x - y$ is a **vector in** $\mathbb{R}^2$, not a scalar.
- Kernels must map to $\mathbb{R}$.
- **Conclusion: Not a valid kernel** (wrong codomain; also not symmetric as a scalar function).

2) $K(x, y) = \sqrt{2}\, x - y$

- Same issue: the output is a **vector**, not a scalar.
- **Conclusion: Not a valid kernel**.

3) $K(x, y) = x^\top (3y)$

- This is $K(x, y) = 3\, x^\top y$. It is scalar and symmetric because $K(x, y) = 3\, x^\top y = 3\, y^\top x = K(y, x)$.
- We need to show Φ such that $\langle \Phi(x), \Phi(y) \rangle = 3\, x^\top y$.

  This holds for

$$\Phi(x) = \sqrt{3}\, x \in \mathbb{R}^2 \quad \Rightarrow \quad \langle \Phi(x), \Phi(y) \rangle = (\sqrt{3}\, x)^\top (\sqrt{3}\, y) = 3\, x^\top y.$$

  Conclusion: Valid kernel, scaled linear kernel with $\Phi(x) = \sqrt{3}\, x$.

Basic concepts solution

↑ View the exercise

Training

1→E, 2→C, 3→D, 4→A, 5→B

Basic math

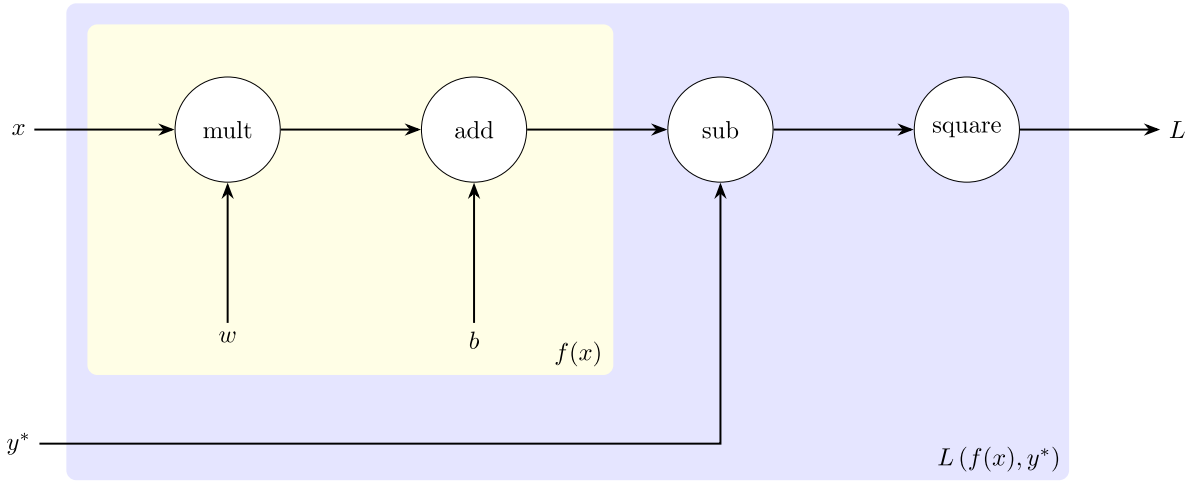
1→C, 2→A, 3→B

Computational graph solution

↑ View the exercise

1) Computational graph

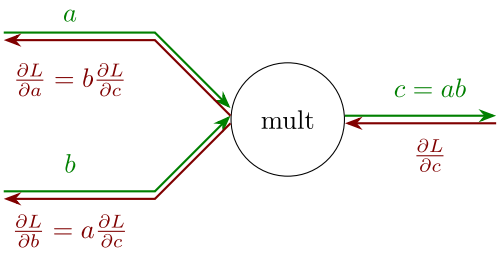
The network can be represented with:

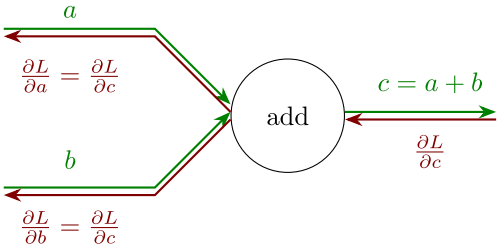


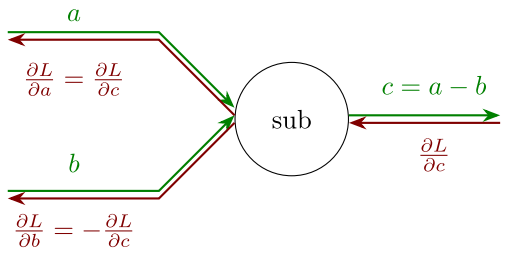
2) Nodes

The forward operation is shown above the arrows in green, the backward operation is shown below the arrows in red.

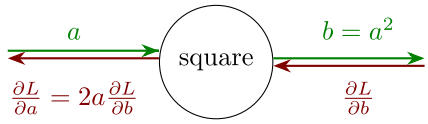
For the multiplication node, we need to store the forward pass inputs to be able to compute the backward pass.







Again, for the square node we need to store the forward pass input.



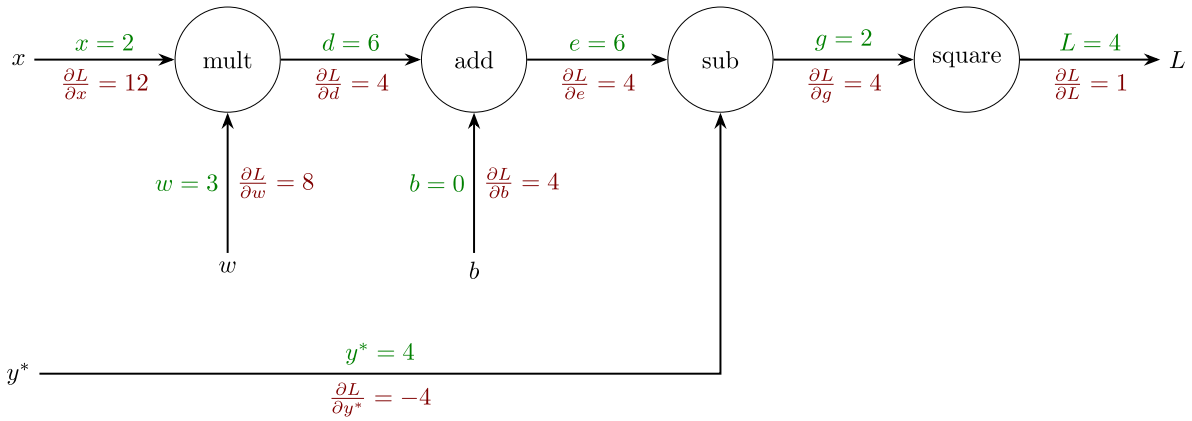
To derive the backward pass we use the chain rule. For example for the square node, we do

$$\frac{\partial L}{\partial a} = \frac{\partial L}{\partial b}\frac{\partial b}{\partial a} = \frac{\partial L}{\partial b}2a$$

3) Full forward pass

Forward pass is shown on top of arrows, in green. The intermediate results are named with arbitrary letters, for example $d = wx$ and $g = (wx + b) - y^*$.

Start on left side and apply the operations (nodes) one by one. The final loss function value is $L = 4$.



4) Full backward pass

Backward pass is shown below the arrows, in red.

Start from the loss and traverse the graph backwards, applying the backward operations of the nodes.

For example to compute the gradient of the loss with respect to the input of the square operation ($\frac{\partial L}{\partial g}$), use the node backward pass

$$\frac{\partial L}{\partial a} = 2a\frac{\partial L}{\partial b}.$$

The gradient incoming from back $\frac{\partial L}{\partial b} = 1$ and the input is named g, so we get $\frac{\partial L}{\partial g} = 2g \cdot 1 = 4$. Note that the b here is the name *inside* the node. Not to be confused with the b in the graph of the whole NN.

The procedure is the same for the rest of the graph.

The final result is

$$\frac{\partial L}{\partial w} = 8, \frac{\partial L}{\partial b} = 4, \frac{\partial L}{\partial x} = 12.$$