Pattern Recognition Exercise Book (B4B33RPZ/BE5B33RPZ)

RPZ team

November 26, 2020

Problems are organized by lecture topics. There are representative problems from tests / exam in previous years, problems recommended for working through lecture / lab materials and more advanced problems.

$\oplus$ – Problems aligned with the lab.
$\odot$ – Problems from previous years tests with solution.
$\star$ – More advanced problem, not to be expected at the exam.

1 Probability, Bayesian Decision Theory

Problem 1.1 (Bayes Theorem)
Prove the Bayes Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

using the axioms of probability:
Axiom 1: $0 \leq P(A) \leq 1$, with $P(A) = 1$ if A is certain.
Axiom 2: If events (A_i), $i = 1, 2, \ldots$ are pairwise incompatible (exclusive) then $P(\bigcup_i A_i) = \sum_i P(A_i)$.
Axiom 3: $P(A \cap B) = P(B|A)P(A)$.

Problem 1.2 (Marginal/Conditional Probabilities)
Consider the same example as in the lecture. The joint probability p_{XK} is given by the table:

	cloudiness			
	1	2	3	4
rain	0.02	0.12	0.09	0.04
no rain	0.38	0.28	0.06	0.01

a) Compute marginal probabilities $p_K(k)$ for $k = \{\text{rain}, \text{no rain}\}$ and $p_X(x)$ for $x = \{1, 2, 3, 4\}$.

b) Compute the probability that the cloudiness is less or equal than 2 given that there was a rain.

Problem 1.3 (Bayes Theorem)
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 had cancer;
- Suppose that 0.1% of the entire population have this disease.

A patient takes a test. Denote the variables as: $C \in \{y, n\}, T \in \{+, -\}$.

a) Compute the probability that a person who test positive has this disease.

b) Compute the probability that a person who test negative does not have this disease.

Problem 1.4 (Coarse Decision Space)
Assume you calculated the posterior probabilities of the state $k \in \{1, \dots 4\}$ as $p_{K|X}(\cdot \mid x) = (0.4, 0.2, 0.2, 0.2)$. The task is to decide whether $k = 1$. What is the optimal Bayesian decision in the following cases (explain):

a) If the penalty for the wrong decision is constant;

b) If mistakenly deciding $k{=}1$ costs twice less than mistakenly deciding that $k{\neq}1$.

Problem 1.5 ($\oplus$ 0-1 loss, 2 classes, Gaussian Conditionals)
Recall the optimal decision strategy q minimizes the risk:

$$R(q) = \sum_{x \in X} \sum_{k \in K} p_{XK}(x, k) W(k, q(x)). \tag{2}$$

Consider 0-1 loss function:

$$W(k, d) = \begin{cases} 1, & \text{if } k \neq d \\ 0, & \text{if } k = d. \end{cases} \tag{3}$$

- Prove: $q(x) = \arg\max_d p(d|x)$.
- Let additionally $K = \{0, 1\}$. Prove $q(x)$ takes the form: $\frac{p(x|k=0)}{p(x|k=1)} \lessgtr \theta$.
- Consider Gaussian Measurements:

$$p(x|k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp(-\frac{(x - \mu_k)^2}{2\sigma_k^2}). \tag{4}$$

  Prove $q(x)$ takes the form $ax^2 + bx + c \lessgtr 0$.

Problem 1.6 (Umbrella Rain)
Consider the setup as in Problem 1.2. You have three possible decisions $D = \{$ umbrella, 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 matrix $W(k, d)$ be as follows:

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

Compute the optimal strategy $q^*(x)$

Problem 1.7 (Error Correcting Codes)
A digital signal transmitting system reads 3 binary digits and for i-th digit outputs the probability that the digit is 1, the resulting probabilities are $0.3, 0.4, 0.7$. It is know that the true digits form an error correcting code where the last digit is always the sum of the first two digits modulo 2.

a) Recognize which number is encoded by the first two digits.

b) Decide whether this packet of 3 digits has to be requested again considering that the cost of skipping an error is 100 more than requesting to repeated the packet.

Problem 1.8 (Gaussian, 3 classes)
We need to classify objects into three classes $k \in \{1, 2, 3\}$. The classes are equally probable a priory. 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 observation 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
$$W = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}.$$
- if the loss matrix is
$$W = \begin{bmatrix} 0 & 2 & 1 \\ 2 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}.$$

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

Problem 1.9 ($\star$ Exam with Bernoulli Chain)
A student prepares for the exam in RPZ. There are K topics in total, one for each lecture. Because the lectures are sequential, he prepares sequentially. He learns the first topic with probability q. If he already learned k topics, he learns the next one again with probability q or otherwise stops preparing.
At the exam he gets a randomly drawn topic. Assume the student answered well on the topic with number x. The task is to recognize whether he/she has prepared at least half of the topics (assume K is even). Model the problem as a Bayesian decision:

a) In this problem, what is the hidden state, observation, decision?

b) What is the probability that he/she learned at least half of the topics?

c) Derive the optimal Bayesian decision strategy.

2 Minimax, Neuman-Pearson, Wald

Problem 2.1 (Student Wants to Marry - *Lecture Example*)
An aging student at CTU wants to marry. He can't 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 $K = \{\text{D}, \text{N}\} \equiv \{\text{F}, \text{M}\}$ (female, male)
- Measurements $X = \{\text{short}, \text{normal}, \text{tall}\} \times \{\text{ultralight}, \text{light}, \text{avg}, \text{heavy}\}$
- Prior probabilities do not exist
- Conditional probabilities $p(x|k)$ are given as follows:

$p(x|\text{F})$

short	.197	.145	.094	.017
normal	.077	.299	.145	.017
tall	.001	.008	.000	.000
	u-light	light	avg	heavy

$p(x|\text{M})$

short	.011	.005	.011	.011
normal	.005	.071	.408	.038
tall	.002	.014	.255	.169
	u-light	light	avg	heavy

(5)

Formulate a Neuman-Pearson strategy for the student: how shall he classify boys and girls?

Problem 2.2 ($\oplus$ Worst Bayes / Minimax)
Consider a binary classification problem ($K{=}\{1,2\}$) with continuous features $x \in \mathbb{R}$. Suppose you have obtained the optimal Bayesian strategy q for the case when the proportion of classes was given by $p(k{=}1) = \pi^*$ (let's call it training distribution prior). Suppose at the test time the proportion of classes changes. What is the worst case performance of the strategy q?

a) How does the risk $R(q) = \sum_k \int p(x,k) W(k, q(x)) dx$ vary as a function of the parameter $\pi = p(k{=}1)$?

b) Show that the maximum over π is achieved either at $\pi = 0$ or at $\pi = 1$.

c) Express the value of the risk in the worst case. Assuming also $0{-}1$ loss, compare the worst risk to the objective of the minimax problem.

d) When the risk $R(q)$ viewed as a function of π is a constant function? Assume this is the case and q is in the form of the likelihood ratio test. Show that q then is the solution to the minimax problem.

Problem 2.3 (Neyman-Pearson for a 2-Class Decision Problem)
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 $p(x|y=1) = 1$ and $p(x|y=2) = x + 0.5$ are given.

a) Write down formally the Neyman-Pearson problem formulation.

b) 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.

Problem 2.4 ($\odot$ Minimax - *Test Example*) Suppose that you have a two-class decision problem $y \in \{1,2\}$ with the real-valued features $x \in [-1,1]$ and that only the class conditional probabilities $p(x|y{=}1) = \max(-x, x) = |x|$ and $p(x|y{=}2) = \min(1+x, 1-x) = 1-|x|$ are given.

a) Write down formally the Minimax problem formulation. Only a mathematical expression will be given points.

b) How many likelihood thresholds are in the solution? Why?

c) Find the optimal Minimax strategy for this decision problem. Any informal solution, e.g. geometric, will be awarded by 0 points.

3 Parameter Estimation, Maximum Likelihood

Problem 3.1 (Bernoulli Coin)
You observed random and independent draws of an unfair coin, the draws were (H, H, H, T, T).

a) Define the probability model and the likelihood of all observations. Find the maximum likelihood estimate of the Heads probability of the coin.

b) Let the heads probability be parametrized as $p = \frac{1}{1+e^{-\eta}}$ for $\eta \in \mathbb{R}$. What is the maximum likelihood estimate of η? Does there always hold $p_{\mathrm{ML}} = \frac{1}{1+e^{-\eta_{\mathrm{ML}}}}$, why?

Problem 3.2 (Binomial Socks)
You have red and blue socks in the drawer in the unknown proportion of red to the total number of socks denoted by p. You draw them randomly with replacement.

a) What does it mean to draw with replacement?

b) What is the probability of drawing two red socks in a row?

c) In $N = 10$ draws you got $R = 2$ red socks. Write the likelihood of such observation. What is the distribution of R for a given N.

d) Compute the maximum likelihood estimate of p.

Problem 3.3 (Gaussian)
The density of a multivariate Normal distribution is given by

$$p(x) = (2\pi)^{-\frac{d}{2}} \det(\Sigma)^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-\frac{1}{2}}(x-\mu)}, \tag{6}$$

where $x, \mu \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d\times d}$, $\Sigma \succcurlyeq 0$. You are given i.i.d. observations $(x_i)_{i=1}^N$.

a) Find the maximum likelihood estimate of the mean μ. *Hint: optimal μ does not depend on Σ.*

b) Find the maximum likelihood estimate of the covariance matrix Σ.
Hint: use log-likelihood; the mean μ is known from above and can be substituted in the end; use the formula $\frac{\partial \det(\Sigma)}{\partial \Sigma} = \Sigma^{-1}\det(\Sigma)$.

Problem 3.4 (Exponential Lamp Lifetime)
At a lamp factory, bulbs are tested in order to know their lifetime. They tested N bulbs and obtained a set of measurements $\mathcal{T} = \{t_1, ..., t_N\}$, where t_i is the time that the light bulb stood before it burned out. We will assume that the measurements follow the exponential distribution:

$$p(t) = \lambda e^{-\lambda t}, \;\; t \in [0, \infty).$$

a) Find the maximum likelihood estimate of the parameter λ.

b) Let us consider an alternative parameterization

$$p(t) = \tfrac{1}{\theta} e^{-\frac{t}{\theta}}.$$

Verify that the expected lifetime of a lamp is θ. Find the maximum likelihood estimate of θ. How does it relate to the rate λ?
Hint: The expected lifetime is the mean of the distribution defined as $\mathbb{E}[t] = \int_0^\infty t p(t) dt$. It can be calculated using integration by parts. An alternative solution is to note that θ_{ML} is the sample mean $\frac{1}{n}\sum_i t_i$ and with $n \to \infty$ the former approaches the true parameter θ and the later the distribution mean.

c) Assume that our prior knowledge about λ is expressed by the distribution $p(\lambda) = e^{-\lambda}$. Derive the MAP estimate of λ, λ_{MAP}. What is the expected lifetime of a bulb for $\hat{\lambda}_{\mathrm{MAP}}$?

Problem 3.5 (Exponential Hard Drives)
The reliability of hard drives is defined by the probability density function $p(t) = \lambda e^{-\lambda t}, t \in (0, \infty)$.

a) What is the maximum likelihood estimate of the failure rate λ if in an experiment with three hard drives the following lifetimes have been observed: $t_1 = 56, t_2 = 120$ and $t_3 = 424$?

b) If the test described above finished at time $T = 300$ and one hard drive would still be running, i.e. $t_1 = 56$ and $t_2 = 120$ are known but about the time of failure of the third hard drive we know only that $t_3 \geq T$. Formulate the likelihood function in this case. Find the ML estimate of λ.

Problem 3.6 ($\star$ German Tank)
During the second world war, British intelligence service had collected information about serial numbers of German tanks ever seen. Suppose serial numbers $x_1, x_2, \ldots x_n$ have been seen. Assume that x_i are independent and follow a uniform distribution with the density

$$p(x) = \frac{1}{\theta}\delta_{\{0 \leq x \leq \theta\}}, \tag{7}$$

where θ is the total number of tanks produced by Germany (we assume $x, \theta \in \mathbb{R}$ for simplicity).

a) What is the maximum likelihood estimate of θ, θ_{ML}?

b) Does the intuition suggest that θ_{ML} underestimates the real number of tanks Germany has?

c) Assume that apriori Germany had capacity to manufacture up to M tanks. What is the maximum a posteriori estimate of θ, θ_{MAP}?

d) Treating θ as a random variables, what is the posterior distribution of θ given the observations. What is the Bayesian estimate of θ minimizing the mean squared error, θ_{MSE}?

4 Nearest Neighbour, Non-Parametric Density Estimation

Problem 4.1 (Piece-wise constant density)
Let $\{x_i\}_{i=1}^n$ be independent observations with $x_i \in [0, 1]$. The domain $[0, 1]$ is partitioned into K equal size segments denoted Δ_k. The piece-wise density model is defined on $[0, 1]$ as

$$p(x) = \begin{cases} d_1, & \text{if } x \in \Delta_1 \\ \ldots & \\ d_K, & \text{if } x \in \Delta_K, \end{cases}$$

where $(d_k \geq 0 \mid |k = 1 \ldots K)$ are parameters.

a) Estimate the parameter vector d using the maximum likelihood.
Hint: use the constraint that the density must integrate to 1.

Problem 4.2 (K-nearest neighbours)
Describe the K-NN algorithm and list its pros and cons.
With the following training set with data points (x, y) (measurement, class), classify point $x = 5$ using 1-NN, 3-NN and 5-NN classifier.

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

Problem 4.3 (K-D trees)
Describe the algorithm for building a K-D tree.
Make a K-D tree (alternating X- and Y- cuts) from the following data:
(2,3), (4,7), (5,4), (7,2), (8,1), (9,6)
Describe how to search for **exact** nearest neighbour using a K-D tree.

Problem 4.4 (Parzen Windows)
Given the measurements $X = \{1, -1, 1, 3, 2, 0\}$, 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:

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

for $h = 2$.

Problem 4.5 (Parzen Window Re-weighting)
Suppose we have training points $\{x_i\}_{i=1}^n$ and found a Parzen density estimate

$$p(y) = \frac{1}{n} \sum_i K(y - x_i)$$

using a fixed kernel K. Here all kernel copies have equal weights $\frac{1}{n}$. Consider giving kernels at different positions a different weight π_i:

$$p(y; \pi) = \sum_j \pi_j K(y - x_j),$$

where $\pi_j \geq 0$ must sum to 1 to ensure p is a density. Re-estimate coefficients π by maximizing the following lower bound on the log likelihood:

$$\sum_i \log p(x_i; \pi) \geq \sum_i \sum_j \frac{K_{i,j}}{K_i} \log \pi_j K_i,$$

where $K_{ij} = K(x_i - x_j)$ and $K_i = \sum_i K_{i,j}$. Later we will see that this is the first iteration of the EM algorithm initialized with $\pi_j = \frac{1}{n}$.

5 Logistic Regression

Problem 5.1 ($\oplus$ Logistic Regression from Bayes Decision)
Consider a recognition problem with two hidden states $K = \{-1, 1\}$ and $x \in \mathbb{R}^d$. We know that the optimal decision expresses in many cases using the likelihood ratio.

a) Assume that the log of likelihood ratio is linear: $\log \frac{p(k=1|x)}{p(k=-1|x)} = w^\mathsf{T} x$. Knowing also that $\sum_k p(k|x) = 1$, find probabilities $p(1|x)$, $p(-1|x)$.

b) Assume that $p(k|x)$ is logistic (as derived in a), $p(x)$ exists but is unknown and does not depend on parameters. Given the training data points $\{x_i, k_i\}_{i=1}^n$, express the negative log likelihood of the data (up to an additive constant).

c) Plot the miss-classification indicator $[\![k \neq \operatorname{sign}(w^\mathsf{T} x)]\!]$ as a function of $z = kw^\mathsf{T} x$. Plot the function $\log(1 + e^{-kw^\mathsf{T} x})$ as a function of same z. Using convexity of this function, show that the negative log likelihood in b) is convex.

Problem 5.2 ($\oplus$ Properties of Logistic Sigmoid Function)
The logistic (sigmoid) function is $\sigma(z) = \frac{1}{1+e^{-z}}$.

a) Show that $\sigma(-z) = 1 - \sigma(z)$;

b) Show that $\frac{\partial}{\partial z}\sigma(z) = \sigma(z)(1 - \sigma(z))$;

c) Compute $\frac{\partial}{\partial z} \log \sigma(z)$ (using b);

d) Compute $\frac{\partial}{\partial x} \log \sigma(w^\mathsf{T} x)$ (using c);

e) Show that $-\log \sigma(z)$ is convex (using monotony of the first derivative or non-negativity of second derivative)

6 Linear Classifier, Perceptron

Problem 6.1 (Linear Classifier)
Let $x \in \mathbb{R}^d$, $a \in \mathbb{R}^d$, $b \in \mathbb{R}$. Consider a linear classifier defined as

$$q(x) = \begin{cases} 1, & \text{if } a^\mathsf{T} x + b \geq 0, \\ -1, & \text{if } a^\mathsf{T} x + b < 0. \end{cases}$$

a) Find the distance from a given point x to the decision boundary of the classifier.
(*Hint: find y on the decision boundary (i.e. satisfying $z^\mathsf{T} y + b = 0$) such that $\|x - y\|^2$ is minimal. Then $d = \|x - y\|$ is the thought distance.*)

b) Let $k_i \in \{1, -1\}$ be the true class of $x_i \in \mathbb{R}^d$ (data point x_i is a vector with coordinates $x_{i,j}$, $j = 1 \dots d$). Define $\bar{x}_i \in \mathbb{R}^{d+1}$ and $w \in \mathbb{R}^{d+1}$ such that q classifies x_i correctly iff $w^\mathsf{T} \bar{x}_i \geq 0$.

Problem 6.2 (Perceptron)
Consider training points $(x_i, k_i)_{i=1}^N$, where $x_i \in \mathbb{R}^d$, $k_i \in \{-1, 1\}$ and let $\bar{x}_i$ be as derived in Problem 6.1 (b). Consider the approximation to the empirical loss:

$$\tilde{R}(w) = \frac{1}{N} \sum_i \max(-w^\mathsf{T} \bar{x}_i, 0), \tag{8}$$

a) Plot this approximation for one data point (x, y) as a function of $z = w^\mathsf{T} \bar{x}$.

b) On the same graph, plot also the empirical loss of one data point as a function of $z = w^\mathsf{T} \bar{x}$.

c) On the same graph, plot also the log likelihood of data point (x, y) in the logistic regression model as a function of z. (see Problem 5.1).

Problem 6.3 (Perceptron)
We will see connection between Perceptron algorithm and stochastic gradient descent for risk approximation (8). Note, the common Perceptron algorithm considers data points sequentially. For a training data as in Problem 6.2, let $l_i(w) = \max(-w^\mathsf{T} \bar{x}_i, 0)$ be the approximate error of one data point and so $\tilde{R}(w) = \frac{1}{N} \sum_i l_i(w)$.

a) Apply stochastic gradient descent to $\tilde{R}$. A step of SGD picks a data point i at random and performs an update

$$w^{t+1} = w^t - \varepsilon \nabla_w l_i(w).$$

(Compute the gradient and simplify what possible).

b) Show that when starting with $w^0 = 0$, the classification boundary at step t is invariant of the step size ε.
(*Hint: inspect what the algorithm would do if starting from w^0 with different values of ε, e.g. $\varepsilon = 1$ and some $\varepsilon \neq 1$*).

Problem 6.4 (Perceptron) A training set is given in the format $T = \{(\mathbf{x_i}; k_i)\}$, where $i = 1 \ldots 5, \mathbf{x_i} \in \mathbb{R}^2$, and $k \in \{1, -1\}$:

$$T = \{(-2, 1; -1), (0, 0; -1), (0, 2; 1), (0, -3; -1), (2, 2, 1)\}.$$

a) Find a linear classifier by the Perceptron algorithm, i.e. find a vector $\mathbf{w} \in \mathbb{R}^2$ and offset $b \in \mathbb{R}$ such that $y = \mathbf{w}\mathbf{x} + b$ is positive for samples from class $k = 1$ end negative for $k = -1$. More specifically, what are the vector $\mathbf{w}$ and offset b after ten steps of the Perceptron algorithm?

Problem 6.5 (Perceptron) For the iterations of the Perceptron algorithm expressed as:

$$w^{t+1} = w^t + x^t,$$

where x^t is the data point selected at step t as misclassified when using weights w^t.

a) Starting with $w^0 = 0$, show that for any t there holds $\|w^t\|^2 \leq t \max_i \|x_i\|^2$.
(*Hint: use induction)*

7 Support Vector Machines

Problem 7.1 (Soft Margin SVM Loss) Consider the primal soft margin SVM formulation:

$$\begin{aligned} \min_{w,b,\xi} \ & \frac{1}{2}\|w\|^2 + C\sum_i \xi_i \\ & (w^\mathsf{T} x_i + b)y_i \geq 1 - \xi_i \quad \forall i, \\ & \xi_i \geq 0 \quad \forall i. \end{aligned}$$

a) Assuming all variables but ξ_i for one i are fixed, find the optimal solution for ξ_i. *Hint: write the respective minimization problem in one variable ξ_i, solve it graphically.*

b) Using the optimal values for ξ_i, reformulate the SVM problem as unconstrained optimization

$$\min_{w,b} \ \frac{1}{2C}\|w\|^2 + \sum_i \max(1 - (w^\mathsf{T} x_i + b)y_i, 0). \tag{9}$$

What optimization methods do you know, that can be applied to solve it?

c) Denoting $z = (w^\mathsf{T} x_i - b)y_i$, plot the function $\max(1 - z, 0)$. C.f. the loss function of Perceptron in Problem 6.2, and logistic regression in Problem 5.1.

d) ($\star$) When $C \to \infty$, does formulation (9) become equivalent to the optimization problem of Perceptron?

e) ($\star$) For a data point i, chosen at random, find the gradient of the function

$$\frac{1}{2nC}\|w\|^2 + \max(1 - (w^\mathsf{T} x_i + b)y_i, 0)$$

in w and b and write a gradient descent step.

Problem 7.2 (Hard Margin SVM) Let $x \in \mathbb{R}^d$, $w \in \mathbb{R}^d$, $b \in \mathbb{R}$. Consider a linear classifier defined as

$$q(x) = \begin{cases} 1, & \text{if } w^\mathsf{T} x + b \geq 0, \\ -1, & \text{if } w^\mathsf{T} x + b < 0. \end{cases}$$

a) The normal vector to the decision hyper-plane is given by $n = \frac{w}{\|w\|}$. Project the vector $x - x^0$ onto n for some x^0 on the hyper-plane (i.e. satisfying $w^\mathsf{T} x^0 + b = 0$).
Hint: Find the scalar product $\langle n, x - x^0 \rangle$ and show that it does not depend on x^0.
Assuming that point x is correctly classified as class $y \in \{-1, 1\}$ show that

$$|\langle n, x - x^0 \rangle| = \frac{(w^\mathsf{T} x + b)y}{\|w\|},$$

called the "signed distance" in the lecture.

b) Let $d_i = \frac{(w^\mathsf{T} x_i + b)y_i}{\|w\|}$ be the signed distance to point x_i. Formulate mathematically the following optimization problem: "find the hyperplane parametrized by w, b that maximizes the smallest absolute distance $|d_i|$ while classifying all points correctly".

c) Show that $\min_i d_i = \max\{d \mid d \leq d_i\}$ for any $d_i \in \mathbb{R}$. Use that to put the problem formulated above in the form

$$\begin{aligned} &\max_{w,b,d} d \\ &\text{s.t.} \quad \frac{(w^\mathsf{T} x_i + b)y_i}{\|w\|} \geq d \ \ \forall i; \quad d \geq 0. \end{aligned}$$

d) Show that the solution is not unique since for any solution (w, b) substituting $(w, b) \mapsto (sw, sb)$ for any $s > 0$ is also a valid solution.

e) Add the constraint $\|w\| = \frac{1}{d}$ and eliminate d.

f) Use that $\frac{1}{\|w\|}$ is a monotone decreasing function of $\|w\|$ to reformulate the problem as minimization of $\|w\|^2$.

Problem 7.3 (SVM: Vector Notation, Dual) Consider the primal soft margin formulation

$$\begin{aligned}\min_{w,b,\xi}\ & \frac{1}{2}\|w\|^2 + C\sum_i \xi_i \\ & (w^\mathsf{T} x_i + b)y_i \geq 1 - \xi_i \quad \forall i, \\ & \xi_i \geq 0 \quad \forall i,\end{aligned}$$

where $w \in \mathbb{R}^d$, $b \in \mathbb{R}$, $(x_i, y_i)_{i=1}^n$ is the training data with $x_i \in \mathbb{R}^d$, $y_i \in \{-1, 1\}$ and $\xi \in \mathbb{R}^n$. Let $\bar{\boldsymbol{X}} \in \mathbb{R}^{d,n}$ be the matrix of all data points multiplied by their class sign, so that $\bar{\boldsymbol{X}}_{:,i} = \boldsymbol{x}_i y_i$. Write the SVM problem using a matrix notation as

$$\begin{aligned}\min_{\boldsymbol{w},b,\boldsymbol{\xi}}\ & \frac{1}{2}\boldsymbol{w}^\mathsf{T}\boldsymbol{w} + C\mathbf{1}^\mathsf{T}\boldsymbol{\xi} \\ & \bar{\boldsymbol{w}}^\mathsf{T}\bar{\boldsymbol{X}} + b\boldsymbol{y}^\mathsf{T} \geq \mathbf{1}^\mathsf{T} - \boldsymbol{\xi}^\mathsf{T} \\ & \boldsymbol{\xi} \geq 0,\end{aligned}$$

where inequalities are coordinate-wise and we used bold fonts to emphasise vectors or matrices (good for books, but not for handwriting).

a) For the constraint $\boldsymbol{w}^\mathsf{T}\bar{\boldsymbol{X}} + b\boldsymbol{y}^\mathsf{T} \geq \mathbf{1}^\mathsf{T} - \boldsymbol{\xi}^\mathsf{T}$ introduce a non-negative vector of Lagrange multipliers $\boldsymbol{\alpha} \in \mathbb{R}^n_+$ to express the constraint as a barrier:

$$\max_{\boldsymbol{\alpha}\geq 0} -(\boldsymbol{w}^\mathsf{T}\bar{\boldsymbol{X}} + b\boldsymbol{y}^\mathsf{T} - \mathbf{1}^\mathsf{T} + \boldsymbol{\xi}^\mathsf{T})\boldsymbol{\alpha} = \begin{cases} 0, & \text{if} \quad \boldsymbol{w}^\mathsf{T}\bar{\boldsymbol{X}} + b\boldsymbol{y}^\mathsf{T} - \mathbf{1}^\mathsf{T} + \boldsymbol{\xi}^\mathsf{T} \geq 0, \\ \infty, & \text{otherwise.} \end{cases}$$

You should obtain the problem reformulation in the form

$$\min_{\boldsymbol{w},b;\boldsymbol{\xi}\geq 0}\ \max_{\boldsymbol{\alpha}\geq 0}(\dots).$$

b) Swap the minimization and maximization and solve analytically for $\boldsymbol{w}, b$ and $\boldsymbol{\xi}$ to obtain maximization in $\boldsymbol{\alpha}$ only.
Hint for $\boldsymbol{w}$: use critical point conditions to obtain that $\boldsymbol{w} = \bar{\boldsymbol{X}}\boldsymbol{\alpha}$.
Hint for $\boldsymbol{\xi}$: use that

$$\min_{\boldsymbol{\xi}\geq 0}(C\mathbf{1} - \boldsymbol{\alpha})^\mathsf{T}\boldsymbol{\xi} = \begin{cases} 0, & \textit{if} \quad C\mathbf{1} - \boldsymbol{\alpha} \geq 0, \\ -\infty, & \textit{otherwise}, \end{cases}$$

in order to eliminate $\boldsymbol{\xi}$ by introducing the constraint $\boldsymbol{\alpha} \leq C$.
Hint for b: use that

$$\min_b b\boldsymbol{y}^\mathsf{T}\boldsymbol{\alpha} = \begin{cases} 0, & \textit{if} \quad \boldsymbol{y}^\mathsf{T}\boldsymbol{\alpha} = 0, \\ -\infty, & \textit{otherwise}, \end{cases}$$

in order to eliminate b by introducing the constraint $\boldsymbol{y}^\mathsf{T}\boldsymbol{\alpha} = 0$.
You should obtain the dual formulation:

$$\begin{aligned}\max_{\boldsymbol{\alpha}}\ & -\frac{1}{2}\boldsymbol{\alpha}^\mathsf{T}\bar{\boldsymbol{X}}^\mathsf{T}\bar{\boldsymbol{X}}\boldsymbol{\alpha} + \mathbf{1}^\mathsf{T}\boldsymbol{\alpha}, \\ & 0 \leq \boldsymbol{\alpha} \leq C, \\ & \boldsymbol{y}^\mathsf{T}\boldsymbol{\alpha} = 0.\end{aligned}$$

Problem 7.4 (Support Vectors) The support vectors are the data points that have an influence on the optimal SVM solution: if coordinates of these points are varied even slightly, the solution $\boldsymbol{w}, b, \boldsymbol{\xi}$ changes.

a) Using the dual relation $\boldsymbol{w} = \bar{\boldsymbol{X}}\boldsymbol{\alpha}$ as derived in Problem 7.3 b), identify which data points can influence $\boldsymbol{w}$ if the solution $\boldsymbol{\alpha}$ stays the same.

b) ($\star$) Argue that if a primal problem constraint (one of the primal inequalities) for a data point $\boldsymbol{x}_i$ was not tight at the optimal solution $\boldsymbol{w}, b, \xi$ (was holding as strict inequality), the same will continue to hold when $\boldsymbol{x}_i$ is varied in a sufficiently small neighborhood and that its Lagrange multiplier α_i will stay zero.

Problem 7.5 (Kernel SVM) Suppose the input features $\boldsymbol{x}_i$ are first lifted to a higher dimension using a lifting function $\phi(\boldsymbol{x})$.

a) Write the dual SVM formulation in this case.

b) Denote the kernel $K(\boldsymbol{x}, \boldsymbol{x}') = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}') \rangle$. Does there always exist a corresponding ϕ for any given mapping $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$?

c) How the classification of the test data can be expressed knowing only the black box computing $K(x, x')$ and not knowing the underlying lifting $\phi(x)$?
Hint: use the support vector form, $\boldsymbol{w} = \bar{\boldsymbol{X}}\boldsymbol{\alpha}$ as derived in Problem 7.3 c.

Problem 7.6 (Kernels and Feature Maps)

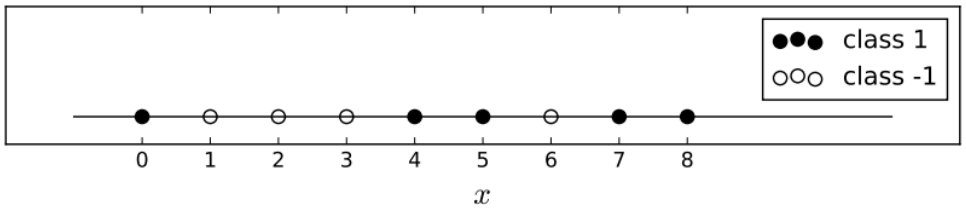
a) Express the kernel function $K(\mathbf{x}, \mathbf{x}')$ that corresponds to feature map $\phi(\mathbf{x}) = (1, x_1, x_2, x_1^2, x_2^2, x_1 \cdot x_2)$ assuming 2-D vectors $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$.

b) Derive the feature map $\phi(\mathbf{x})$ that corresponds to kernel function $K(\mathbf{x}, \mathbf{x}') = (1 + x \cdot x')^2$ for $\mathbf{x} \in \mathbb{R}^2$.

c) Let us have a kernel function $K(\mathbf{x}, \mathbf{x}') = (1 + x \cdot x')^d$ and $\mathbf{x} \in \mathbb{R}^D$ for $d, D \in \mathbb{N}$. Do we know how to compute feature map for given d and D? Compare the computation of the explicit feature map and the kernel function.

8 Adaboost

Problem 8.1 (1-D Adaboost Classifier)
Adaboost learning algorithm. Consider the following 1-D data:



and the following set of weak classifiers: $h(x) = \text{sign}(ax + b)$ $(a, b \in R)$. Use this example to explain how Adaboost works (make one full iteration, ending with first data re-weighting.)

9 Neural Networks, Backpropagation

10 K-Means Clustering

11 Principal Component Analysis, Fisher Linear Discriminant

12 Decision Trees

13 EM algorithm

ANSWERS

Problem 1.2

a) $p_K(\text{rain}) = 0.27$, $p_X(1) = 0.4$.
b) $P(x \leq 2|\text{rain}) \approx 0.52$.

Problem 1.3

a) $P_{C|T}(y|+) \approx 3\%$.

Problem 1.4

a) The optimal decision is "no".

Problem 2.4

a)

We define the objective function of the minimax task as

$$\arg\min_{q:X\to Y} \max_{y\in Y} \sum_{x:q(x)\neq y} p(x \mid y), \tag{10}$$

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

b)

For a 2-class 2-decision Minimax problem there is always only one likelihood threshold. This comes from the fact that the decision $q^*(x) = d^*$ is the solution of the system of inequalities

$$\gamma(x)c_1(d^*) + c_2(d^*) \leq \gamma(x)c_1(d) + c_2(d), \qquad d \in D \setminus \{d^*\}, \tag{11}$$

where c_1, c_2 are constants, and $\gamma(x)$ is the likelihood ratio. The system is linear with respect to the likelihood ratio. Therefore, in case of two decisions, e.g. $D = \{1, 2\}$, there will be only one threshold.

c)

First, we plot the probability distributions and the likelihood ratio:

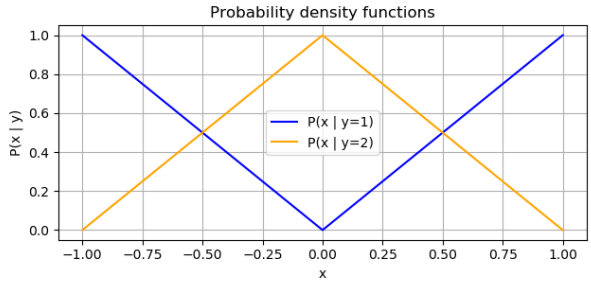


Fig. 1.a: Probability density functions.

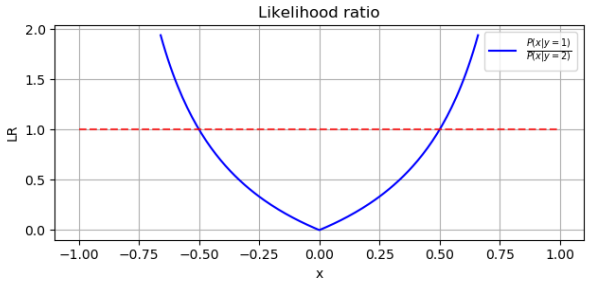


Fig. 1.b: Likelihood ratio $\gamma(x) = \frac{p(x|y=1)}{p(x|y=2)}$.

The minimax task

$$\arg\min_{q:X\to Y} \max_{y\in Y} \sum_{x:q(x)\neq y} p(x \mid y), \tag{12}$$

can be rewritten as

$$\arg\min_{q(x)} \max \left\{ \int_{X_2} p(x \mid y=1)\, dx, \int_{X_1} p(x \mid y=2)\, dx \right\} \tag{13}$$

where

$$X_1 \subseteq X, x \in X_1 : q(x) = 1, \tag{14}$$
$$X_2 \subseteq X, x \in X_2 : q(x) = 2, \tag{15}$$
$$X_1 \cup X_2 = X, \tag{16}$$
$$X_1 \cap X_2 = \emptyset. \tag{17}$$

We are looking for $t \in \langle -1, 1\rangle$, where

$$X_1 = \langle -1, -t\rangle \cup \langle t, 1\rangle, \tag{18}$$
$$X_2 = \langle -t, t\rangle. \tag{19}$$

Therefore the task is now

$$\arg\min_{q(x)} \max \left\{ \int_{-t}^{t} p(x \mid y=1)\, dx, \int_{-1}^{-t} p(x \mid y=2)\, dx + \int_{t}^{1} p(x \mid y=2)\, dx \right\} \tag{20}$$

and thanks to the likelihood ratio $\gamma(x)$ being symmetrical around $x = 0$

$$= \arg\min_{q(x)} \max \left\{ \int_{-t}^{0} p(x \mid y=1)\, dx, \int_{-1}^{-t} p(x \mid y=2) \right\}. \tag{21}$$

If exists $q(x)$ such that

$$\int_{-t}^{0} p(x \mid y=1)\, dx = \int_{-1}^{-t} p(x \mid y=2)\, dx \tag{22}$$

then $-t, t$ are points where $q(x)$ changes and $q(x)$ is the optimal strategy.

$$\int_{-t}^{0} -x\,dx = \int_{-1}^{-t} 1 + x\,dx \tag{23}$$

$$-\left[\frac{x^2}{2}\right]_{-t}^{0} = \left[x + \frac{x^2}{2}\right]_{-1}^{-t} \tag{24}$$

$$\frac{t^2}{2} = \frac{t^2}{2} - t + 1 - \frac{1}{2} \tag{25}$$

$$t = \frac{1}{2} \tag{26}$$

$$X_1 = \left\langle -1, -\frac{1}{2} \right\rangle \cup \left\langle \frac{1}{2}, 1 \right\rangle \tag{27}$$

$$X_2 = \left\langle -\frac{1}{2}, \frac{1}{2} \right\rangle \tag{28}$$

Now, we compute the likelihood ratio threshold θ. Since we compute the threshold by substituting to the functions for $x \in (-1, 0)$, we substitute the x in both $p(x \mid y = 1)$ and $p(x \mid y = 2)$ for the value of $-t = -\frac{1}{2}$.

$$\theta = \frac{-x}{1+x} = \frac{\frac{1}{2}}{1 - \frac{1}{2}} = 1. \tag{29}$$

Finally, we define the optimal strategy

$$q(x) = \begin{cases} 1, & \text{if } \gamma(x) \geq \theta \\ 2, & \text{else.} \end{cases} \tag{30}$$

Problem 6.1

a) The distance is $d = \frac{|a^\mathsf{T} x + b|}{\|a\|}$.