Machine Learning 1 — Repeat Final Exam

1 Preliminaries

- Please write your immatriculation number **but not your name** on *every* page you hand in.
- The exam is closed book. You may, however, take one A4 sheet of handwritten notes.
- The exam is limited to 2×60 minutes.
- If a question says "Describe in 2–3 sentences" or "Show your work" or something similar, these mean the same: give a succinct description or explanation.
- This exam consists of 9 pages, 15 problems. You can earn up to 44 points. With two points less you can get a perfect grade on this exam.

Problem 1 [2 points] We want to reduce our gender bias in grading. To help us, please fill in your immatriculation number on every sheet you hand in. Make sure it is easily readable. But then, make sure you do **not** write your name on *any* sheet you hand in.

Can you do that, please? Thanks. You can earn two extra points if you do so.

2 Probability

Problem 2 [2 points] A Munich weather forecast app can forecast 4 kinds of weather—rainy, sunny, snowy and cloudy. The accuracy of rainy forecast is 0.8, while the accuracy of sunny, snowy and cloudy forecast is 0.9. In the past 5 years, Munich had 10 percent rainy days. If the app shows that tomorrow is a rainy day, what is the probability that it is not going to rain?

We use A and W to replace the app and the real weather.

$$
\begin{aligned}
p(A = rain \mid W = rain) &= 0.8 \\
p(A = notrain \mid W = rain) &= 0.2 \\
p(A = rain \mid W = notrain) &= 0.1 \\
p(A = notrain \mid W = notrain) &= 0.9 \\
p(W = rain) &= 0.1 \\
p(W = notrain) &= 0.9 \\
p(W = notrain \mid A = rain) & \\
= \frac{p(A = rain \mid W = notrain)p(W = notrain)}{p(A = rain \mid W = notrain)p(W = notrain) + p(A = rain \mid W = rain)p(W = rain)} & \\
= \frac{0.1 \times 0.9}{0.1 \times 0.9 + 0.8 \times 0.1} &= 0.53
\end{aligned}
$$

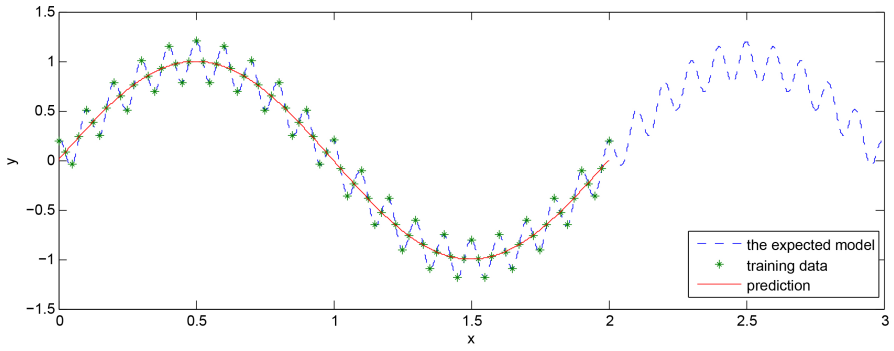
3 Neural networks

We have data with input $\mathbf{x} \in \mathbb{R}$ and output $\mathbf{y} \in \mathbb{R}$ (see the Figure). The training data is generated from $y = \sin(\pi x) + 0.2\cos(20\pi x)$. We use neural networks with one input, one output and 40 hidden units to approximate the data. The cost function is

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} \|z(x_n, \mathbf{w}) - y_n\|^2 + \lambda \mathbf{w}^T \mathbf{w}$$

where $z(x_n, \mathbf{w})$ is the prediction of x_n.

The activation function that is used on the hidden units only is $\phi(x) = \tanh(x)$, while the single output unit is linear.



Problem 3 [2 points] What is the reason that the model ignores the information of $0.2\cos(20\pi x)$? It is known that the size of the training data set is large enough.

The regularisation is a low-pass filter.

Problem 4 [1 point] If the input training data is in the range of $[0, 2]$, plot the prediction in the input data range $(2, 4]$.

It cannot get a good result out of the training input range.

Problem 5 [2 points] If we use a linear activation for the hidden units, what would the result be? Show your work.

It is linear.

4 Coin

Problem 6 [4 points] You have two coins, C_1 and C_2. Let the outcome of a coin toss be either *heads* ($C_i = 1$) or *tails* ($C_i = 0$) for $i = 1, 2$. C_1 is a fair coin, i.e., it has an equal prior on *heads* and *tails*. However, C_2 depends on C_1: If C_1 shows *heads* ($C_1 = 1$), C_2 will show *heads* with probability 0.7. If C_1 shows *tails* ($C_1 = 0$), C_2 will show *heads* with probability 0.5. Now you toss C_1 and C_2 in

sequence once. You observe the sum of the two coins $S = C_1 + C_2 = 1$. What is the probability that C_1 shows *tails* and C_2 shows *heads*? *(Hint: Bayes' rule.)*

We know:

$$\begin{aligned} P(C_1 = H = 1) = P(C_1 = T = 0) &= 0.5 \\ P(C_2 = H = 1 | C_1 = 1) &= 0.7 \\ P(C_2 = H = 1 | C_1 = 0) &= 0.5 \\ S = C_1 + C_2 &= 1 \end{aligned}$$

We want to know:

$$P(C_1 = 0, C_2 = 1 | S = 1)$$

Therefore, we need Bayes rule [1 point]:

$$\begin{aligned} P(C_1, C_2 | S) &\overset{\text{Bayes}}{=} \frac{P(S|C_1, C_2)P(C_1, C_2)}{\text{norm. const.}} \\ &\overset{\text{chain rule}}{=} \frac{P(S|C_1, C_2)P(C_1)P(C_2|C_1)}{\text{norm. const.}} \\ &\overset{\text{expand}}{=} \frac{P(S|C_1, C_2)P(C_1)P(C_2|C_1)}{\sum_{C_1', C_2'} P(S|C_1', C_2')P(C_1')P(C_2'|C_1')} \end{aligned}$$

Solve for asked probability [1 point]:

$$\Rightarrow P(C_1 = 0, C_2 = 1 | S = 1) = \frac{\overbrace{P(S = 1 | C_1 = 0, C_2 = 1)}^{=1} \overbrace{P(C_1 = 0)}^{=0.5} \overbrace{P(C_2 = 1 | C_1 = 0)}^{=0.5}}{\sum_{C_1', C_2'} P(S = 1 | C_1', C_2')P(C_1')P(C_2'|C_1')}$$

Expand denominator [1 point]:

$$\begin{aligned} \sum_{C_1', C_2'} P(S = 1 | C_1', C_2')P(C_1')P(C_2'|C_1') &= \underbrace{P(S = 1 | C_1' = 0, C_2' = 0)}_{=0} P(C_1' = 0)P(C_2' = 0 | C_1' = 0) \\ &+ \underbrace{P(S = 1 | C_1' = 1, C_2' = 0)}_{=1} P(C_1' = 1)P(C_2' = 0 | C_1' = 1) \\ &+ \underbrace{P(S = 1 | C_1' = 0, C_2' = 1)}_{=1} P(C_1' = 0)P(C_2' = 1 | C_1' = 0) \\ &+ \underbrace{P(S = 1 | C_1' = 1, C_2' = 1)}_{=0} P(C_1' = 1)P(C_2' = 1 | C_1' = 1) \\ &= P(C_1' = 1)P(C_2' = 0 | C_1' = 1) + P(C_1' = 0)P(C_2' = 1 | C_1' = 0) \\ &= 0.5 \times (1 - 0.7) + 0.5 \times 0.5 \\ &= 0.15 + 0.25 = 0.4 \end{aligned}$$

Write down final answer [1 point]:

$$\begin{aligned} P(C_1 = 0, C_2 = 1 | S = 1) &= 0.25/0.4 \\ &= 0.625 = \frac{5}{8} \end{aligned}$$

5 Linear Regression

You want to boost your Facebook page and therefore you book Facebook advertisements. A simple linear model for the number of new likes per week (y), depending on the money spent (x) could be:

$$\begin{aligned} y &= a_0 + a_1 x + \epsilon \\ \text{where } y &= \text{number of new likes per week} \\ x &= \text{money spent in that week, in units of 1 EUR} \\ \epsilon &= \text{normal (Gaussian) distributed fluctuations} \end{aligned}$$

After taking a lot of measurement data you fit the parameters. You find:

$$\begin{aligned} a_0 &= 10 \\ a_1 &= 5 \\ \mathrm{E}[y] &= 0 \\ \mathrm{var}[y] &= 4 \end{aligned}$$

The full model is therefore given by

$$\begin{aligned} y &= 10 + 5x + \mathcal{N}(0, 4) \\ &= 10 + 5x + (8\pi)^{-1/2} \exp(-x^2/8) \end{aligned}$$

Problem 7 [3 points] Assume you spend no money, what is the probability that you get more than 10 likes per week?

$$\begin{aligned} x &= 0 \\ \Rightarrow y &= 10 + \epsilon \longrightarrow \mathcal{N}(10, 4) \\ \Rightarrow p(y > 10) &= \int_{10}^{\infty} \mathcal{N}(10, 4) dy \\ &= 0.5 \end{aligned}$$

Problem 8 [3 points] Now you spend 1 EUR on advertisements. What is the expected value of likes?

$$\begin{aligned} x &= 1 \\ \Rightarrow y &= 10 + 5 + \epsilon \longrightarrow \mathcal{N}(15, 4) \\ \Rightarrow \mathrm{E}[y] &= 15 \end{aligned}$$

6 Multivariate Normal

Consider a bivariate Gaussian distribution $p(x_1, x_2) = \mathcal{N}(\mathbf{x} \mid \mu, \mathbf{\Sigma})$ where

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$$

Problem 9 [3 points] Compute $p(x_2 \mid x_1)$ for the case $\sigma_1 = \sigma_2 = 1$ and $\sigma_{12} = \sigma_{21} = \rho$. Remember that

$$\begin{aligned} p(x_2 \mid x_1) &= \mathcal{N}(x_2 \mid \mu_{2|1}, \Sigma_{2|1}) \\ \mu_{2|1} &= \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1) \\ \Sigma_{2|1} &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \end{aligned}$$

$$\begin{aligned} \mu_{2|1} &= \mu_2 + \rho(x_1 - \mu_1) \\ \Sigma_{2|1} &= 1 - \rho^2 \\ \implies p(x_2 \mid x_1) &= \mathcal{N}(x_2 \mid \mu_2 + \rho(x_1 - \mu_1), 1 - \rho^2) \end{aligned}$$

Problem 10 [3 points] Give a graphical interpretation for the conditional obtained in the previous problem.

$p(x_2 \mid x_1)$ is obtained by "slicing" the joint distribution through the $X_1 = x_1$ line (cf. figure 4.9 on p. 112, Murphy).

7 Logistic Regression

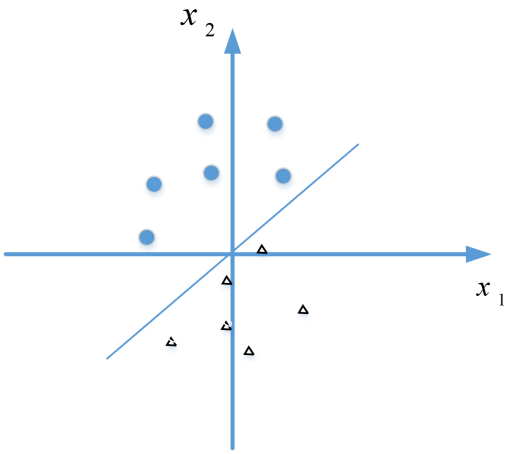
We employ a logistic regression model to classify the data which are plotted in the below figure,

$$\mathbf{P}(Y = 1 \mid \mathbf{x}, w_1, w_2) = \frac{1}{1 + \exp(-w_1 x_1 - w_2 x_2)}.$$

We fit the data by the maximum likelihood approach, and minimise

$$J(\mathbf{w}) = -l(\mathbf{w}).$$

We get the decision boundary as shown in the figure, and the error of the classification is 0.

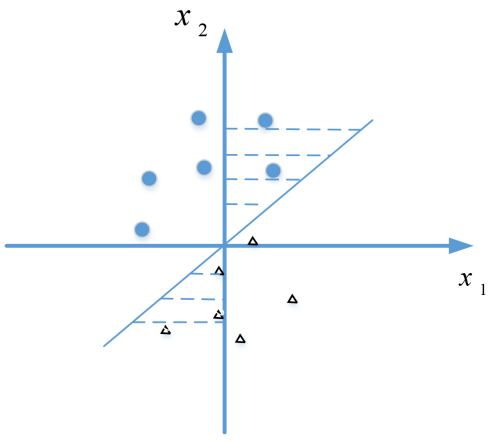


Problem 11 [3 points] Now, we regularise w_2 and minimise

$$J_0(\mathbf{w}) = -l(\mathbf{w}) + \lambda w_2^2.$$

Draw the area that the decision boundary can be and explain your work.

When we regularise w_2, the decision boundary becomes more vertical. If λ is extremely large, the decision boundary is x_2 axis.



8 Kernels

The following informations about kernels *might* be helpful for solving the next two problems.

Let K_1 and K_2 be kernels on $\mathcal{X} \subseteq \mathbb{R}^n$, then the following functions are kernels:

1. $K(\boldsymbol{x}, \boldsymbol{y}) = K_1(\boldsymbol{x}, \boldsymbol{y}) + K_2(\boldsymbol{x}, \boldsymbol{y})$
2. $K(\boldsymbol{x}, \boldsymbol{y}) = \alpha\, K_1(\boldsymbol{x}, \boldsymbol{y})$ for $\alpha > 0$
3. $K(\boldsymbol{x}, \boldsymbol{y}) = K_1(\boldsymbol{x}, \boldsymbol{y})\, K_2(\boldsymbol{x}, \boldsymbol{y})$
4. $K(\boldsymbol{x}, \boldsymbol{y}) = K_3(\boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{y}))$ for K_3 kernel on $\mathbb{R}^m$ and $\boldsymbol{\phi} : \mathcal{X} \to \mathbb{R}^M$
5. $K(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^T B \boldsymbol{y}$ for $B \in \mathbb{R}^{n \times n}$ symmetric and positive semi-definite

The following identities involving the exponential function *might* be helpful for solving the next two problems.

$$\exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$
$$\exp(x) = \lim_{n \to \infty} \left(1 + \frac{x}{n}\right)^n$$
$$\exp(a + b) = \exp(a) \exp(b)$$
$$\exp(a\, b) = \exp(a)^b$$

Problem 12 [6 points] Let Z be a set of *finite* size. Show that the function

$$K_0(X, Y) = |X \cap Y|$$

is a valid kernel, provided that $X \subseteq Z$ and $Y \subseteq Z$. Remember that Z is finite, i.e. $Z = \{z_1, z_2, \ldots, z_N\}$.

Enumerate all elements of Z, i.e. $Z = \{z_1, z_2, \ldots, z_N\}$. This is possible because Z is of finite cardinality.

Define the feature map $\phi : 2^Z \to \mathbb{R}^N$ by

$$\phi_i(X) = \begin{cases} 1 & \text{if } z_i \in X \\ 0 & \text{if } z_i \notin X \end{cases}.$$

We have

$$K_0(X, Y) = \sum_{i=1}^{N} \underbrace{\phi_i(X)\phi_i(Y)}_{\begin{cases} = 1 & \text{if } z_i \in X \wedge z_i \in Y \\ = 0 & \text{otherwise} \end{cases}} = |X \cap Y|\,.$$

Problem 13 [4 points] Again, let Z be a set of *finite* size. Show that the function

$$K(X, Y) = 2^{|X \cap Y|}$$

is a valid kernel, provided that $X \subseteq Z$ and $Y \subseteq Z$.

Even if you did not succeed in the previous exercise, you may assume that $K_0(X, Y)$ is a valid kernel.

Set

$$K_1(X, Y) = (\log 2) K_0(X, Y) .$$

This is a kernel (multiplication of kernel by positive constant).

The Taylor expansion of the exponential function is

$$\exp(K_1(\boldsymbol{x}, \boldsymbol{y})) = 1 + \sum_{n=1}^{\infty} \frac{1}{n!} K_1(\boldsymbol{x}, \boldsymbol{y})^n .$$

The power $K_1(\boldsymbol{x}, \boldsymbol{y})^n$ is a kernel by iterated application of rule 3 ($K_1(\boldsymbol{x}, \boldsymbol{y})K_2(\boldsymbol{x}, \boldsymbol{y})$ is a kernel). The product $(1/n!)K_1(\boldsymbol{x}, \boldsymbol{y})^n$ is a kernel by rule 2 ($\alpha K_1(\boldsymbol{x}, \boldsymbol{y})$ if a kernel for $\alpha > 0$) because $(1/n!)$ is always positive. The sum $\sum_{n=1}^{\infty} 1/(n!)K_1(\boldsymbol{x}, \boldsymbol{y})^n$ is a kernel by iterated application of rule 1 ($K_1(\boldsymbol{x}, \boldsymbol{y}) + K_2(\boldsymbol{x}, \boldsymbol{y})$ is a kernel). The constant 1 is a kernel by rule 4 ($K_3(\boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{y}))$) with $K_3(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^T \boldsymbol{y}$ and $\phi(\boldsymbol{z}) = (1)$. Thus $1 + \sum_{n=1}^{\infty} \frac{1}{n!} K_1(\boldsymbol{x}, \boldsymbol{y})^n$ is a kernel by rule 1. This was previously shown in lecture as part of the exercise "The Gaussian kernel" (tutor session about kernels) and you may thus use this result without reproofing it.

Thus

$$K(X, Y) = \exp((\log 2)|X \cap Y|) = \exp(\log 2)^{|X \cap Y|} = 2^{|X \cap Y|}$$

is a valid kernel.

9 Constrained Optimisation

Suppose we have 40 pieces of raw material. Toy A can be made of one piece material with 3 Euro machining fee. A larger toy B can be made from two pieces of material with 5 Euro machining fee.

We can sell x pieces of toy A for $20 - x$ Euro each, and y pieces of toy B for $40 - y$ Euro each.

From our experience, toy B is more popular than toy A; therefore, we will produce not more of toy A than of toy B. To get the maximum profit, we want to calculate the number toy A and toy B that we should produce.

Problem 14 [3 points] Write down the problem using the primal optimisation method.

Problem 15 [3 points] The problem can be solved using Karush–Kuhn–Tucker (KKT) conditions. Write down these conditions (but don't solve them).

$$\begin{aligned} \min f(x, y) = &- (\ x(20 - x) + y(40 - y) - 3x - 5y) = x^2 - 17x + y^2 - 35y \\ s.t.\ & x + 2y \leq 40 \\ & x \leq y \end{aligned}$$

$$L(x, y, \alpha_1, \alpha_2) = x^2 - 17x + y^2 - 35y + \alpha_1(x + 2y - 40) + \alpha_2(x - y)$$

$$\frac{\partial L}{\partial x} = 2x - 17 + \alpha_1 + \alpha_2 = 0$$
$$\frac{\partial L}{\partial y} = 2y - 35 + 2\alpha_1 - \alpha_2 = 0$$
$$\alpha_1(x + 2y - 40) = 0$$
$$\alpha_2(x - y) = 0$$
$$x + 2y \leq 40$$
$$x \leq y$$
$$\alpha_1, \alpha_2 \geq 0$$