Machine Learning 1 — 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 8 pages, 15 problems. You can earn up to 44 points.

Problem 1 [3 points] Fill in your immatriculation number on every sheet you hand in. Make sure it is easily readable. Make sure you do **not** write your name on *any* sheet you hand in.

2 Linear Algebra and Probability Theory

Problem 2 [2 points] Let X and Y be two random variables. Show that

$$\mathrm{var}[X+Y] = \mathrm{var}[X] + \mathrm{var}[Y] + 2\mathrm{cov}[X,Y]$$

where cov[X,Y] is the covariance between X and Y. You can use that

$$\mathrm{var}[X] = \mathrm{E}[X^2] - \mathrm{E}^2[X]$$
$$\mathrm{cov}[X,Y] = \mathrm{E}[XY] - \mathrm{E}[X]E[Y]$$

We know [1 point]:

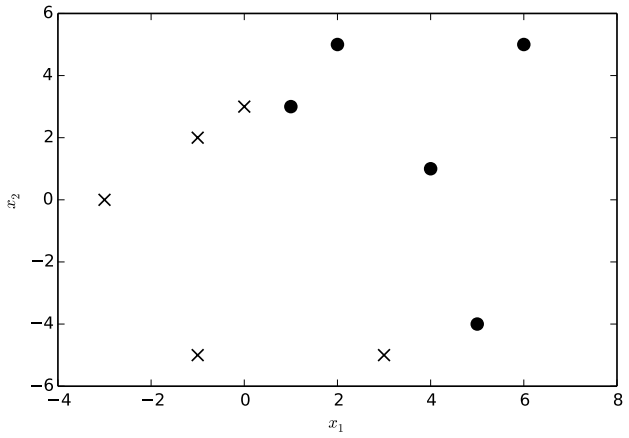
$$\mathrm{var}[X] = \mathrm{E}[X^2] - \mathrm{E}^2[X]$$
$$\mathrm{cov}[X,Y] = \mathrm{E}[XY] - \mathrm{E}[X]E[Y]$$

Hence [1 point]:

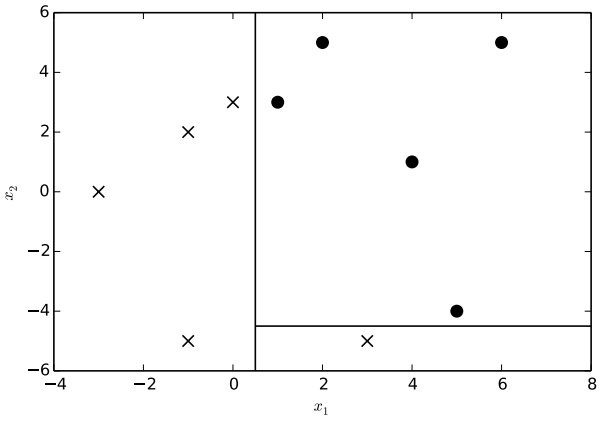
$$\begin{aligned}
\mathrm{var}[X+Y] &= \mathrm{E}[(X+Y)^2] - \mathrm{E}^2[X+Y] \\
&= \mathrm{E}[X^2 + Y^2 + 2XY] - (\mathrm{E}[X] + \mathrm{E}[Y])^2 \\
&= \mathrm{E}[X^2] + \mathrm{E}[Y^2] + 2\mathrm{E}[XY] - \mathrm{E}^2[X] - \mathrm{E}^2[Y] - 2\mathrm{E}[X]\mathrm{E}[Y] \\
&= \underbrace{\mathrm{E}[X^2] - \mathrm{E}^2[X]}_{=\mathrm{var}[X]} + \underbrace{\mathrm{E}[Y^2] - \mathrm{E}^2[Y]}_{=\mathrm{var}[Y]} + 2\underbrace{(\mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y])}_{=\mathrm{cov}[X,Y]} \qquad \square
\end{aligned}$$

3 Decision Trees & kNN

You are given two-dimensional input data with corresponding targets in below plot.



Problem 3 [2 points] Sketch the decision boundaries of a maximally-trained decision tree classifier using misclassification rate.

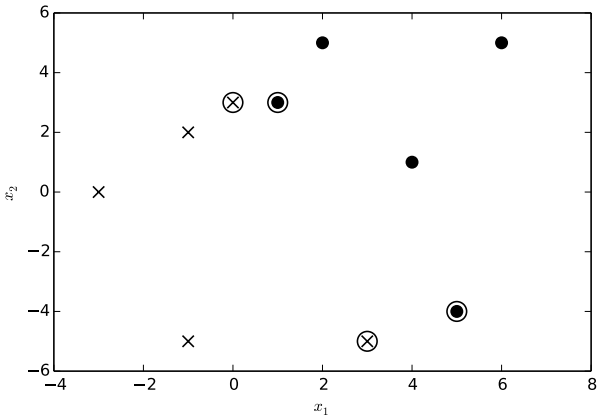


Problem 4 [2 points] Describe how this model can overfit the data. Describe how that problem can be solved or prevented.

Overfitting!

- prune the tree
- restrict the depth of the tree
- train an ensemble of slightly different trees –> random forests

Problem 5 [2 points] Perform 1-NN with leave-one-out cross validation on the data in the plot. Circle all points that are misclassified and write down the accuracy.



accuracy = 6/10

4 Linear Classification

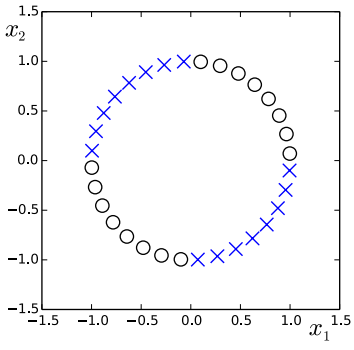
Problem 6 [4 points] The decision boundary for some linear classifier on two-dimensional data crosses axis x_1 at 2 and x_2 at 5. First, write down the general form of a linear classifier model (how many parameters do you need, given the dimensions?). Calculate the coefficients (parameters).

$$
\begin{aligned}
w_0 + w_1 x_1 + w_2 x_2 &= 0 \\
w_0 + 2w_1 &= 0 \\
w_1 &= -\frac{w_0}{2} \\
w_0 + 5w_2 &= 0 \\
w_2 &= -\frac{w_0}{5}
\end{aligned}
$$

$$
\begin{aligned}
\text{set, e.g., } w_0 &= -2 \\
w_1 &= 1 \\
w_2 &= \frac{2}{5}
\end{aligned}
$$

or anything proportional

Problem 7 [2 points] Which basis function $\phi(x_1, x_2)$ makes the data in the example below linearly separable (crosses in one class, circles in the other)?



$$\phi(x_1, x_2) = x_1 x_2$$

5 Gaussian Processes

The posterior conditional for an MVN with distribution

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix} \sim \mathcal{N}\left(\mu = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}\right)$$

is given by

$$\begin{aligned} p(\boldsymbol{y}_1|\boldsymbol{y}_2) &= \mathcal{N}(\boldsymbol{y}_1|\boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2}) \\ \boldsymbol{\mu}_{1|2} &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{y}_2 - \boldsymbol{\mu}_2) \\ \boldsymbol{\Sigma}_{1|2} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \end{aligned}$$

Assume a noise-free GP with mean function

$$m(x) = 0$$

and covariance function

$$K(x, x') = 1 + (x - 2)(x' - 2).$$

You are given two data points $(0, 4) =: \boldsymbol{x}$.

Problem 8 [2 points] Compute the kernel matrix (aka covariance matrix) for $\boldsymbol{x}$.

$$\boldsymbol{K} = \begin{pmatrix} 5 & -3 \\ -3 & 5 \end{pmatrix}$$

Problem 9 [6 points] Given corresponding outputs $\boldsymbol{y} = (2, 6)$, compute the posterior function values for data points $\boldsymbol{x}_* = (0, 2, 4)$.

$$
\begin{aligned}
\boldsymbol{K}^{-1} &= \begin{pmatrix} \frac{5}{16} & \frac{3}{16} \\ \frac{3}{16} & \frac{5}{16} \end{pmatrix} \\
\boldsymbol{K}_*^T &= \begin{pmatrix} 5 & -3 \\ 1 & 1 \\ -3 & 5 \end{pmatrix} \\
\mu_{\boldsymbol{f}_*|\boldsymbol{y}} &= m(\boldsymbol{x}_*) + \boldsymbol{K}_*^T \boldsymbol{K}^{-1}(\boldsymbol{y} - m(\boldsymbol{x})) \\
&= 0 + \begin{pmatrix} 1 & 0 \\ 1/2 & 1/2 \\ 0 & 1 \end{pmatrix} \left(\begin{pmatrix} 2 \\ 6 \end{pmatrix} - 0 \right) \\
&= \begin{pmatrix} 2 \\ 4 \\ 6 \end{pmatrix}
\end{aligned}
$$

Problem 10 [4 points] Which other algorithm does this resemble? Please describe the corresponding feature space.

K is a linear kernel, and we are basically solving linear regression.

6 Neural networks

Problem 11 [2 points] Geoffrey has a data set with input $X \in \mathbb{R}^2$ and output $Y \in \mathbb{R}^1$. He has neural network A with one hidden layer and 9 neurons in that layer, which can fit the data. However, he does not know how good the model is, so he also tests neural network B with two hidden layers and three neurons for each of these layers. Both models have biases for the hidden units only.

- How many free parameters do the two models have? Show your calculation, not just the result.
- What are the pros and cons of model A compared to model B? Mention at least one pro and one con.

Model A and model B have 36 and 24 free parameters, respectively. Pros: fewer parameters for B. Cons: more complicated structure may prone to local minimum.

Problem 12 [4 points] You know that the sum of squared errors is related to the Gaussian distribution—differently put, if you assume a normal distribution of the data around their expectation, the maximum likelihood estimate (MLE) is reached when the summed squared errors is minimised.

The same is true for a Laplace distribution and the sum of absolute errors. In particular, if the data observes a Laplacian distribution

$$
p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{w}, \beta) = \prod_{n=1}^{N} p(z_n|x_n, \boldsymbol{w}, \beta) = \prod_{n=1}^{N} \frac{1}{2\beta} \exp\left(-\frac{|z_n - y(x_n, \boldsymbol{w})|}{\beta} \right)
$$

then minimising the summed absolute errors

$$\sum_{n=1}^{N} |z_n - y(x_n, \boldsymbol{w})|$$

leads to MLE. In these equations, $\boldsymbol{x}$ is the vector of all inputs, x_n is the input of sample n, while $y(x_n, \boldsymbol{w})$ is the neural network prediction on x_n. Then, z_n is the desired output for x_n.

Show that the MLE of the Laplace distribution minimises the sum of absolut errors.

Taking the negative logarithm, we obtain the error function

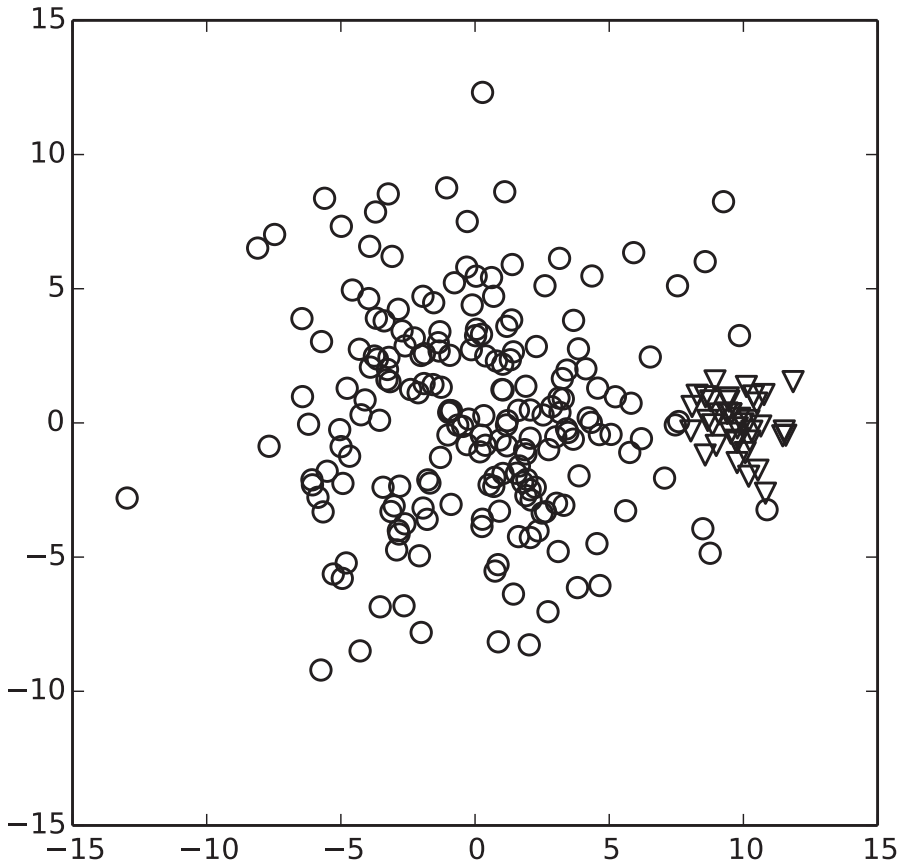
$$\frac{\beta}{2} \sum_{n=1}^{N} \big|z_n - y(x_n, \boldsymbol{w})\big| - \frac{N}{2} \ln \beta + \frac{N}{2} \ln(2\pi).$$

Maximising the likelihood function is equivalent to minimising the error given by the sum of absolutes

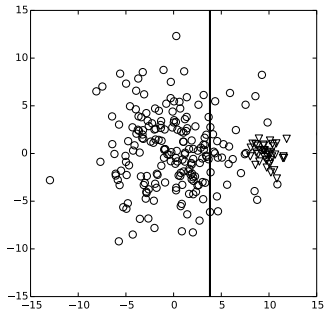
$$\sum_{n=1}^{N} |z_n - y(x_n, \boldsymbol{w})|.$$

7 Unsupervised learning

Problem 13 [2 points] Consider the plot below. The two classes are circles and triangles. What significance do class labels have for k-means? Draw the resulting decision boundary in the plot for k-means with two centroids.

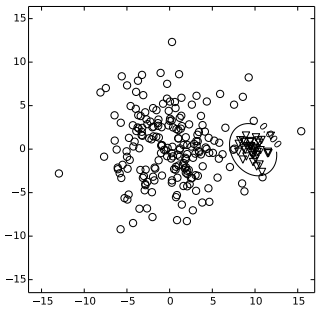


Exact position is not important, but should not seperate the two classes perfectly.



Problem 14 [3 points] Would the separation be different using the EM algorithm with a Gaussian mixture model using two components and individual full covariance matrices? What would it look like? The left cluster has 200 points and the right cluster has 40 points. Draw qualitatively in the above figure.

EM for GMM would be able to distinguish better between the two classes. k-means only fits clusters of the same size. EM for GMM can model different sizes for each cluster. The new boundary is now a circle around the right cluster:



Problem 15 [4 points] The likelihood for ICA is

$$\ell(\boldsymbol{W}) = \sum_{i=1}^{m} \sum_{j=1}^{n} \log p_{\boldsymbol{S}_i}(\boldsymbol{w}_i^T \boldsymbol{x}) + \log |\boldsymbol{W}|$$

When calculating the gradient for this likelihood we need to compute the derivative of $\log p_{\boldsymbol{S}_i}(\boldsymbol{w}_i^T \boldsymbol{x})$. Here we are using $p_{\boldsymbol{S}_i}(s) \approx \sigma'(s)$ where the sigmoid function is given as

$$\sigma(a) = \frac{1}{1 + e^{-a}}.$$

with the derivative

$$\sigma'(a) \equiv \frac{d\sigma(a)}{da} = \sigma(a)\left(1 - \sigma(a)\right).$$

Show that $\frac{d}{ds} \log \sigma'(s) = 1 - 2\sigma(s)$.

$\frac{d}{ds} \log \sigma'(s) = \frac{\sigma''(s)}{\sigma'(s)} = \frac{\frac{d}{ds} \frac{e^{-s}}{(1+e^{-s})^2}}{\frac{e^{-s}}{(1+e^{-s})^2}} = \frac{\frac{e^{-s}(e^{-s}-1)}{(1+e^{-s})^3}}{\frac{e^{-s}}{(1+e^{-s})^2}} = \frac{e^{-s}-1}{1+e^{-s}} = \frac{e^{-s}+1-2}{1+e^{-s}} = 1 - \frac{2}{1+e^{-s}} = 1 - 2\sigma(s)$