Machine Learning — Final Exam

1	2	3	4	5	6	7	8	9	10	11	12	$\sum$
5	7	6	7	5	6	8	7	3	6	9	8	**??**

Do not write anything above this line

Name:

Student ID: Signature:

- Only write on the sheets given to you by supervisors. If you need more paper, ask the supervisors.
- Pages 15-18 can be used as scratch paper.
- All sheets (including scratch paper) have to be returned at the end.
- **Do not unstaple the sheets!**
- Wherever answer boxes are provided, please write your answers in them.
- Please write your student ID (*Matrikelnummer*) on every sheet you hand in.
- **Only use a black or a blue pen (no pencils, red or green pens!).**
- You are allowed to use your A4 sheet of handwritten notes (two sides). **No other materials (e.g. books, cell phones, calculators) are allowed!**
- Exam duration - 120 minutes.
- This exam consists of **??** pages, **??** problems. You can earn **??** points.

1 Decision Trees

Problem 1 [(1+4)=5 points] You are developing a model to classify games at which machine learning will beat the world champion within five years. The following table contains the data you have collected.

x_1 (Team or Individual)	x_2 (Mental or Physical)	x_3 (Skill or Chance)	y (Win or Lose)
T	M	S	W
I	M	S	W
T	P	S	W
I	M	C	W
T	P	S	L
I	M	C	L
T	P	C	L
T	P	C	L
T	P	C	L
I	P	S	W

You can look up the value of $\log_2(x)$ in this table:

x	0.10	0.2	0.25	0.33	0.50	0.66	0.75	0.8	1.0
$\log_2(x)$	-3.32	-2.32	-2.0	-1.60	-1.0	-0.60	-0.42	-0.32	0.0

a) Calculate the entropy $i_H(y)$ of the class labels y.

$$p(y = W) = p(y = L) = 1/2$$
$$H(y) = p(y = W) \log p(y = W) + p(y = L) \log p(y = L) = 1$$

b) Build the optimal decision tree of depth 1 using entropy as the impurity measure. Which attribute is selected as the root of the decision tree?

We need to compute the information gain Δ for splitting on each of the features x_1, x_2 and x_3, and pick the feature with the highest information gain.

Splitting on x_1:

$$p(y = W|x_1 = T) = 2/6 \qquad p(y = W|x_1 = I) = 3/4$$
$$i_H(x_1 = T) = 1/3 \cdot \log(1/3) + 2/3 \cdot \log(2/3) \approx 0.93$$
$$i_H(x_1 = I) = 3/4 \cdot \log(3/4) + 1/4 \cdot \log(1/4) = 0.815$$
$$\Delta(x_1) = 1 - \frac{6}{10} \cdot 0.93 - \frac{4}{10} \cdot 0.815 \approx 0.12$$

Splitting on x_2:

$$p(y = W|x_2 = M) = 1/4 \qquad p(y = W|x_2 = P) = 2/6$$
$$\Delta(x_2) = \Delta(x_1)$$

Splitting on x_3:

$$p(y = W|x_3 = S) = 4/5 \qquad p(y = W|x_3 = C) = 1/5$$
$$i_H(x_3 = S) = 4/5 \cdot \log(4/5) + 1/5 \cdot \log(1/5) \approx \frac{3.6}{5}$$
$$\Delta(x_3) = 1 - \frac{1}{2}\frac{3.6}{5} - \frac{1}{2}\frac{3.6}{5} = 1 - \frac{3.6}{5} = \frac{1.4}{5} = 0.28$$

We would split on x_3 since it yields the highest information gain.

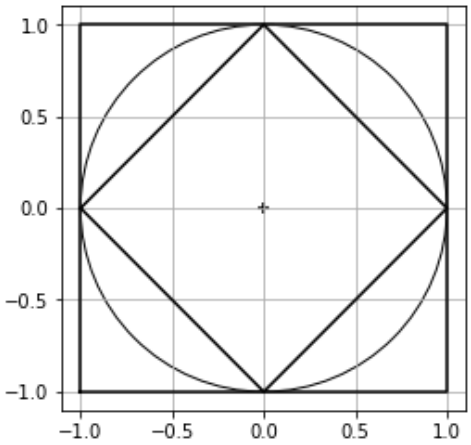
(HW1, problem 1)

2 KNN

Problem 2 [(1.5+5.5)=7 points]

a) Let $\boldsymbol{x} \in \mathbb{R}^2$. Draw the unit circle for the following norms. Make sure to clearly label which circle corresponds to which norm.

- L_1-norm: $||\boldsymbol{x}||_1 = \sum_i |x_i|$
- L_2-norm: $||\boldsymbol{x}||_2 = \sqrt{\sum_i x_i^2}$
- L_∞-norm: $||\boldsymbol{x}||_\infty = \max_i |x_i|$



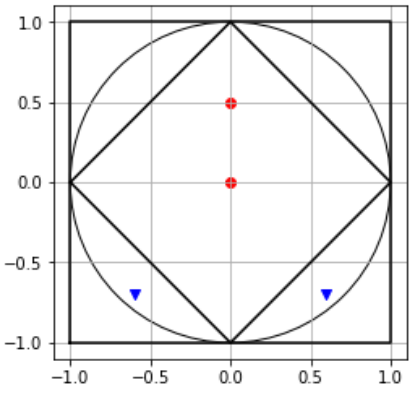
(Lecture 1, slide 13)

b) Construct a binary classification dataset that consists of 4 data points, that is specify $\boldsymbol{x}_i \in \mathbb{R}^2$ and $y_i \in \{0, 1\}$ (i.e. write the coordinates and labels) for each data point i, such that:

> Performing leave-one-out cross validation (LOOCV) with a 1-NN (one nearest neighbor) classifier using L_1 distance yields 0% misclassification rate. Meanwhile, performing LOOCV with a 1-NN classifier using L_∞ distance on the same dataset yields misclassification rate of 50%.

Hint: Remember the shape of the unit circles.

Class red: $p_1 = (0, 0)$, $p_2 = (0, 0.5)$ Class blue: $q_1 = (-0.6, -0.7)$, $q_2 = (0.6, -0.7)$



$$L_1(p_1, q_1) = 1.3, \qquad L_1(p_1, q_2) = 1.3, \qquad L_1(q_1, q_2) = 1.2$$
$$L_\infty(p_1, q_1) = 0.7, \qquad L_\infty(p_1, q_2) = 0.7, \qquad L_\infty(q_1, q_2) = 1.4$$

(Mock exam, Problem 2)

3 Probabilistic Inference

Problem 3 [(6)=6 points] A kangaroo starts from a random location $\boldsymbol{x}_0 \in \mathbb{R}^2$ in the jungle and after one jump reaches the location $\boldsymbol{x}_1 \in \mathbb{R}^2$.

The prior over the start location $\boldsymbol{x}_0$ is the standard bivariate normal distribution

$$p(\boldsymbol{x}_0) = \mathcal{N}\left(\boldsymbol{x}_0 \;\middle|\; \begin{bmatrix}0\\0\end{bmatrix}, \boldsymbol{I}_{2\times 2}\right),$$

where $\boldsymbol{I}_{2\times 2}$ denotes the 2 by 2 identity matrix.

The conditional distribution $p(\boldsymbol{x}_1|\boldsymbol{x}_0)$ is a normal distribution with mean $\boldsymbol{x}_0$ and identity covariance

$$p(\boldsymbol{x}_1|\boldsymbol{x}_0) = \mathcal{N}\big(\boldsymbol{x}_1|\boldsymbol{x}_0, \boldsymbol{I}_{2\times 2}\big)$$

Assume that we observe $\boldsymbol{x}_1$, the position of the kangaroo after the jump.

Write down the closed-form expression for $p(\boldsymbol{x}_0|\boldsymbol{x}_1)$. Make sure that you obtain a valid probability distribution (i.e. it integrates to one). Show your work.

Important: You are not allowed to simply use facts about conditionals of multivariate normal distribution (e.g. from Bishop's book). Derive the result starting from the Bayes formula.

Start with the Bayes formula

$$\begin{aligned}
p(\boldsymbol{x}_0|\boldsymbol{x}_1) &\propto p(\boldsymbol{x}_1|\boldsymbol{x}_0)p(\boldsymbol{x}_0) \\
&\propto \mathcal{N}\left(\boldsymbol{x}_0 \;\middle|\; \begin{bmatrix}0\\0\end{bmatrix}, \boldsymbol{I}_{2\times 2}\right) \times \mathcal{N}\big(\boldsymbol{x}_1|\boldsymbol{x}_0, \boldsymbol{I}_{2\times 2}\big) \\
&\propto \exp\left(-\frac{1}{2}\boldsymbol{x}_0^T\boldsymbol{x}_0\right) \exp\left(-\frac{1}{2}(\boldsymbol{x}_0^T - \boldsymbol{x}_1)^T(\boldsymbol{x}_0 - \boldsymbol{x}_1)\right)
\end{aligned}$$

Simplify & absorb all the terms constant in $\boldsymbol{x}_0$ into $\propto$

$$\propto \exp\left(-\frac{1}{2}\left(2\boldsymbol{x}_0^T\boldsymbol{x}_0 - 2\boldsymbol{x}_1^T\boldsymbol{x}_0\right)\right)$$

We see that there is a quadratic term $2\boldsymbol{x}_0^T\boldsymbol{x}_0$ and a linear term $-2\boldsymbol{x}_1^T\boldsymbol{x}_0$ in the exponent. Therefore we conclude that this is a normal distribution.

$$\begin{aligned}
&\propto \mathcal{N}(\boldsymbol{x}_0 \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&\propto \exp\left(-\frac{1}{2}(\boldsymbol{x}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_0 - 2\boldsymbol{x}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\right)
\end{aligned}$$

We need to complete the square and determine its parameters $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$.

$$\begin{aligned}
\boldsymbol{x}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_0 &\stackrel{!}{=} 2\boldsymbol{x}_0^T\boldsymbol{x}_0 \iff \boldsymbol{\Sigma} = \frac{1}{2}\boldsymbol{I} \\
2\boldsymbol{x}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} &\stackrel{!}{=} 2\boldsymbol{x}_0^T\boldsymbol{x}_1 \iff \boldsymbol{\mu} = \frac{1}{2}\boldsymbol{x}_1
\end{aligned}$$

Therefore the posterior distribution is

$$p(\boldsymbol{x}_0 \mid \boldsymbol{x}_1) = \mathcal{N}\left(\boldsymbol{x}_0 \mid \frac{1}{2}\boldsymbol{x}_1, \frac{1}{2}\boldsymbol{I}\right)$$

(HW3, Problem 4, HW12, Problem 4)

4 Regression

Problem 4 [(7)=7 points] Consider the following one-dimensional regression problem:

$$p(y_i \mid w, x_i, \tau) = \mathcal{N}(y_i \mid wx_i, \tau^2)$$
$$p(w \mid \sigma) = \mathcal{N}(w \mid 0, \sigma^2)$$

where $x_i, y_i \in \mathbb{R}$, and $\tau > 0$ and $\sigma > 0$ are variance parameters. You fit the parameter w using, e.g., the MLE or the MAP approach on a dataset $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_N)\}$ of N i.i.d. instances.

Your task is to qualitatively describe what happens to the quantities in each column as the parameters in each row vary. Specifically, in each cell of the table below you have to write one and only one of the following three options: *(i) increases*, *(ii) decreases* or *(iii) no change*.

For example in the top left cell you have to specify whether the quantity $\mathrm{Var}(p(w|\sigma))$ increases, decreases or does not change as we increase the value of the parameter σ.

	$\mathrm{Var}(p(w\|\sigma))$	$\lvert w_{MLE} - w_{MAP}\rvert$	$\lvert\mathbb{E}_{p(w\|\mathcal{D})}[w] - w_{MAP}\rvert$
σ increases	increases	decreases	no change
σ decreases	decreases	increases	no change
N increases	no change	decreases	no change

In the table above

- $\mathrm{Var}(p(w|\sigma))$ denotes the variance of the distribution $p(w|\sigma)$
- w_{MLE} and w_{MAP} denote the maximum likelihood estimate and the MAP estimate of the parameter w respectively.
- $\mathbb{E}_{p(w|\mathcal{D})}[w]$ is the expectation of the posterior distribution over w.
- $|\cdot|$ denotes the absolute value.

Explanation:

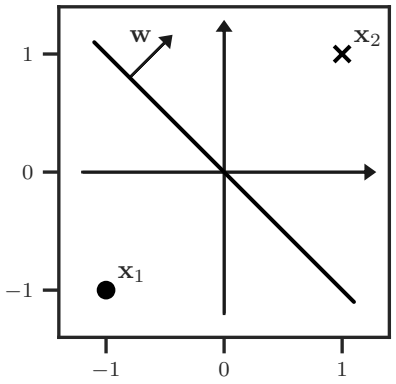
- The prior $p(w|\sigma^2)$ is a normal distribution $\mathcal{N}(w|0, \sigma^2)$, variance is equal to $\mathrm{Var}(p(w|\sigma)) = \sigma^2$ *(basic properties of normal distribution)*. Increasing σ increases the variance (and vice versa). Changing the number of samples N has no effect on the prior $p(w|\sigma)$.
- A prior with larger variance is less informative (more flat), i.e. has less influence on the posterior. *(Lecture 3, slide 59)*

As we observe more samples (N increases), the effect of the prior diminishes and w_{MAP} approaches w_{MLE}. *(Lecture 2, slide 42)*

- Since the posterior distribution $p(w|\mathcal{D})$ is a Gaussian, the mode w_{MAP} is always equal to the mean $\mathbb{E}_{p(w|\mathcal{D})}[w]$ *(Lecture 3, slide 59)*.

5 Classification

Problem 5 [(5)=5 points] Consider the classification problem in the figure below. There are two points, $\mathbf{x}_1 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$ with class $y_1 = 0$ and $\boldsymbol{x}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ with class $y_2 = 1$. Further you are given a logistic regression model with weight vector $\boldsymbol{w} = \begin{bmatrix} \log 2 \\ \log 2 \end{bmatrix}$. Here, log denotes natural logarithms, i.e. base e. Assume that there is no bias term.



Prove or disprove that the weight vector $\boldsymbol{w}$ is the MAP estimate for a logistic regression model with a Gaussian prior on $\boldsymbol{w}$ with precision $\lambda = 1$.

MAP estimate for $\boldsymbol{w}$ is defined as

$$\begin{aligned} \boldsymbol{w}_{MAP} &= \arg\max_{\boldsymbol{w}} p(\boldsymbol{w}|\mathcal{D}) \\ &= \arg\min_{\boldsymbol{w}} -\log p(\boldsymbol{w}|\mathcal{D}) \end{aligned}$$

$\boldsymbol{w}_{MAP}$ is the argmin if the gradient is zero, since the loss function is convex and differentiable *(HW5, Problem 4)*

$$\Leftrightarrow \nabla_{\boldsymbol{w}} -\log p(\boldsymbol{w}_{MAP}|\mathcal{D}) \stackrel{!}{=} 0$$

Objective function:

$$\begin{aligned} \mathcal{L}(\boldsymbol{w}) &= -\log p(\boldsymbol{w}|\mathcal{D}) \\ &= -\log p(\mathcal{D}|\boldsymbol{w}) - \log p(\boldsymbol{w}|\lambda) \\ &= -\sum_{i=1}^{2} y_i \log \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i) + (1-y_i)\log(1-\sigma(\boldsymbol{w}^\top \mathbf{x}_i)) + \frac{\lambda}{2}\mathbf{w}^\top \mathbf{w} + \text{const.} \end{aligned}$$

(Lecture 4, slides 35-37)

Compute the gradient of the negative log posterior.

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = -\sum_{i=1}^{n} \boldsymbol{x}_i \left[y_i - \sigma\left(\boldsymbol{w}^\top \boldsymbol{x}_i\right) \right] + \lambda \boldsymbol{w}.$$

(HW4, Problem 5, Task 3)

Plugging in our numbers we get

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} &= -\boldsymbol{x}_1 \left[0 - \sigma\left(\boldsymbol{w}^\top \boldsymbol{x}_1\right)\right] - \boldsymbol{x}_2 \left[1 - \sigma\left(\boldsymbol{w}^\top \boldsymbol{x}_2\right)\right] + \lambda \boldsymbol{w} \\ &= -\begin{bmatrix}1/5\\1/5\end{bmatrix} - \begin{bmatrix}1/5\\1/5\end{bmatrix} + \begin{bmatrix}\log 2\\\log 2\end{bmatrix} = \begin{bmatrix}-2/5 + \log 2\\-2/5 + \log 2\end{bmatrix}\end{aligned}$$

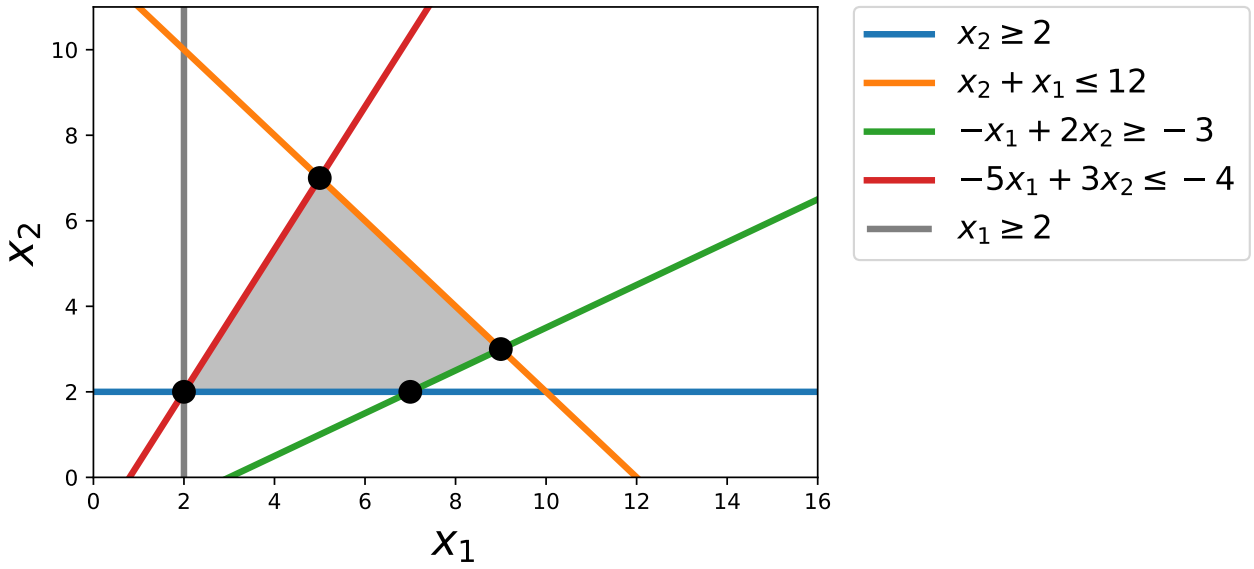
Since the gradient w.r.t. the weights is not zero the solution is not optimal. That is, $\boldsymbol{w}$ is not the MAP estimate of the logistic regression model.

6 Constrained Optimization

Problem 6 [(3+3)=6 points] Consider the following optimization problem

$$\begin{aligned} \min_{x_1,x_2} \; & 2x_1 - 3x_2 \\ x_1 &\geq 2 \\ x_2 &\geq 2 \\ x_1 + x_2 &\leq 12 \\ -x_1 + 2x_2 &\geq -3 \\ -5x_1 + 3x_2 &\leq -4 \end{aligned}$$

a) Draw the set of feasible points



(HW6, Problem 2a)

b) Solve the optimization problem, i.e. find the minimizer (x_1^*, x_2^*).

The domain (feasible set) is a convex set. Minimizing a concave function is equivalent to maximizing a convex function. We know from the lecture that in this case the optimum lies on one of the vertices of the domain. That is, we only need to check the points $(2, 2)$, $(7, 2)$, $(5, 7)$ and $(9, 3)$. By checking these 4 points we conclude that $(x_1^*, x_2^*) = (5, 7)$ is the solution.

(Lecture 6, slides 13-14)

7 SVM

Problem 7 [(7+1)=8 points]

a) Consider training a hard-margin SVM using a linear kernel $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^T \boldsymbol{x}_j$ on a linearly separable training set $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \dots (\boldsymbol{x}_N, y_N)\}$. Let s denote the number of support vectors we would obtain if we would train on the entire dataset. Furthermore, let ε denote the leave-one-out cross validation (LOOCV) misclassification rate.

Does the following inequality hold? Justify your answer.

$$\varepsilon \leq \frac{s}{N}$$

Intuitively, the result follows from the following claim (which we prove below): if $\boldsymbol{x}_i$ is not a support vector when training on the entire training set, then the optimal $\boldsymbol{w}^*$ and b^* do not change when leaving $\boldsymbol{x}_i$ out of the training set.

Since the original data are linearly separable and since we are using a hard-margin classifier, the hypothesis given by the original $\boldsymbol{w}^*$ and b^* will not make an error on $\boldsymbol{x}_i$, and hence, no error will be made in the i-th step of the LOOCV. Equivalently, the only *possible* errors in the LOOCV procedure are made on $\boldsymbol{x}_i$'s which are support vectors when training on the entire training set, and hence $\varepsilon \leq \frac{s}{N}$.

(The following details are not required for full points).

Formally, let $(\boldsymbol{w}^*_{\mathcal{D}}, b^*_{\mathcal{D}})$ and $\boldsymbol{\alpha}^*_{\mathcal{D}}$ denote the optimal primal and dual solutions for the SVM when training on the entire $\mathcal{D}$. Also let, $\mathcal{D}_i = \mathcal{D} \backslash \{(\boldsymbol{x}_i, y_i)\}$ be the set of training examples when omitting the i-th example, and let $(\boldsymbol{w}^*_{\mathcal{D}_i}, b^*_{\mathcal{D}_i})$ and $\boldsymbol{\alpha}^*_{\mathcal{D}_i}$ be the optimal primal and dual variables of the optimization problem when training on $\mathcal{D}_i$.

$\boldsymbol{\alpha}_{\mathcal{D}_i}$ consists of only $n-1$ variables, namely $\boldsymbol{\alpha}_{\mathcal{D}_i,1}, \dots, \boldsymbol{\alpha}_{\mathcal{D}_i,i-1}, \boldsymbol{\alpha}_{\mathcal{D}_i,i+1}, \dots, \boldsymbol{\alpha}_{\mathcal{D}_i,N}$. Now consider setting the dual variables as follows $\boldsymbol{\alpha}_{\mathcal{D}_i,j} = \boldsymbol{\alpha}^*_{\mathcal{D},j}$ for $j \neq i$. Note that, if $\boldsymbol{x}_i$ is not a support vector when training on $\mathcal{D}$ then $\boldsymbol{\alpha}^*_{\mathcal{D},i} = 0$. We can verify that $(\boldsymbol{w}^*_{\mathcal{D}}, b^*_{\mathcal{D}})$ and $\boldsymbol{\alpha}_{\mathcal{D}_i}$ satisfy the KKT conditions for the SVM optimization problem when training on $\mathcal{D}_i$ (e.g. the condition regarding the derivatives of the Lagrangian with respect to the primal variables is guaranteed to hold by our construction). From this, and the fact that $\boldsymbol{w}^*_{\mathcal{D}}, b^*_{\mathcal{D}}$ are unique since the objective function is strictly convex we can conclude that $\boldsymbol{w}^*, b^*$ do not change when omitting $\{(\boldsymbol{x}_i, y_i)\}$ as desired.

b) Consider a setting similar to the previous problem, except that we now we use SVM with an arbitrary valid kernel k. Assume that the data is linearly separable in the feature space corresponding to the kernel. Does $\varepsilon \leq \frac{s}{N}$ hold in this case? Justify your answer.

Yes. The above argument only uses the facts that the optimum of a convex optimization problem is not affected by leaving out non-active constraints, and that the training data can be perfectly classified by the obtained hypothesis based on training on the full dataset. The choice of kernel has no influence.

Note that the support vectors may, of course, be different when we are using different kernels.

8 Kernels

Problem 8 [(7)=7 points] Let $\mathcal{M}$ denote the set of all real-valued matrices of arbitrary size.

Prove or disprove that the function $k : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ is a valid kernel.

$$k(X, Y) = \min\{\text{rank}(X), \text{rank}(Y)\}$$

A kernel $k : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ is valid, if there exists a feature map $\phi : \mathcal{M} \to \mathcal{H}$, such that $k(X, Y) = \langle \phi(X), \phi(Y) \rangle_{\mathcal{H}}$, where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is the inner product on $\mathcal{H}$.

Consider the feature map $\phi(X) = (\underbrace{1, 1, 1, ..., 1}_{\text{rank}(X) \text{ ones}}, \underbrace{0, 0, 0, 0, 0, 0, 0, 0...}_{\text{zeros everywhere else}})$.

Then $\phi(X)^T \phi(Y) = \sum_{j=1}^{\infty} \phi_j(X)\phi_j(Y) = \min\{\text{rank}(X), \text{rank}(Y)\} = k(X, Y)$ is the inner product. Hence, $k(X, Y)$ is a valid kernel.

(HW 7, Problems 4, 5, 6)

9 Deep Learning

Problem 9 [(1+1+1)=3 points] You are given a dataset containing images $\boldsymbol{x}_i \in [0, 1]^D$ (all the pixel values are normalized between 0 and 1) and respective class labels $y_i \in \{1, ..., C\}$. You implement a fully connected neural network with two hidden layers, *tanh* activations and L_2 regularization on all weights excluding biases.

a) Consider two strategies for initializing the weights of your neural network.

   1) Sample the weights from Uniform$(-10, 10)$

   2) Sample the weights from Uniform$(-1, 1)$

   Which choice (1 or 2) is more reasonable, given that we are training the network with backpropagation? Justify your answer.

   Large magnitudes of the weights may lead to saturation of the *tanh* activation and vanishing gradients. Therefore we choose option (2).

b) When training neural networks, why do we usually stop training when the loss on the validation set starts to increase?

   Overfitting.

c) After training has finished, your model has high training loss and high validation loss. What should you do? Justify your answer.

   Increase the number of parameters (width or depth of network), lower regularization.

10 Dimensionality Reduction

Problem 10 [(6)=6 points] Let the matrix $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ represent N data points of dimension $D = 10$ (samples stored as rows). We applied PCA to $\boldsymbol{X}$. By using the $K = 5$ top principal components, we transformed/projected $\boldsymbol{X}$ into $\tilde{\boldsymbol{X}} \in \mathbb{R}^{N \times K}$. We computed that $\tilde{\boldsymbol{X}}$ preserves 70% of the variance of the original data $\boldsymbol{X}$.

Suppose now we apply PCA on the following matrices:

a) $\boldsymbol{Y}_1 = \boldsymbol{X}\boldsymbol{S}$ where $\boldsymbol{S} = \lambda \boldsymbol{I}$, with $\lambda \in \mathbb{R}$ and $\boldsymbol{I} \in \mathbb{R}^{D \times D}$ is the identity matrix

b) $\boldsymbol{Y}_2 = \boldsymbol{X}\boldsymbol{R}$ where $\boldsymbol{R} \in \mathbb{R}^{D \times D}$ and $\boldsymbol{R}\boldsymbol{R}^T = \boldsymbol{I}$

c) $\boldsymbol{Y}_3 = \boldsymbol{X}\boldsymbol{P}$ where $\boldsymbol{P} = \text{diag}(+5, -5, \ldots, +5, -5)$ is a $D \times D$ diagonal matrix

d) $\boldsymbol{Y}_4 = \boldsymbol{X}\boldsymbol{Q}$ where $\boldsymbol{Q} = \text{diag}(1, 2, 3, \ldots, D-1, D)$ is a $D \times D$ diagonal matrix

e) $\boldsymbol{Y}_5 = \boldsymbol{X} + \boldsymbol{1}_N \boldsymbol{\mu}^T$ where $\boldsymbol{\mu} \in \mathbb{R}^D$ and $\boldsymbol{1}_N$ is an N-dimensional column vector of all ones

f) $\boldsymbol{Y}_6 = \boldsymbol{X}\boldsymbol{A}$ where $\boldsymbol{A} \in \mathbb{R}^{D \times D}$ and $\text{rank}(\boldsymbol{A}) = 5$

and obtain the projected data $\tilde{\boldsymbol{Y}}_1, \ldots \tilde{\boldsymbol{Y}}_6 \in \mathbb{R}^{N \times K}$ using the principal components corresponding to the top $K = 5$ largest eigenvalues of the respective $\boldsymbol{Y}_i$.

What fraction of variance of each $\boldsymbol{Y}_i$ will be preserved by each respective $\tilde{\boldsymbol{Y}}_i$? Justify your answer.

The answer "cannot tell without additional information" is also valid if you provide a justification.

Let the eigendecomposition of the covariance matrix $\boldsymbol{X}^T\boldsymbol{X}$ be

$$\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T.$$

The fraction of variance preserved is defined as

$$\frac{\sum_{i=1}^{K} \lambda_i}{\sum_{i=1}^{D} \lambda_i}$$

where $\lambda_1 \geq ... \geq \lambda_D$ are the eigenvalues (diagonal elements of $\boldsymbol{\Lambda}$) sorted in descending order.

a) 70%. All eigenvalues are scaled by the same amount λ^2, so the fraction doesn't change.

b) 70%. $\boldsymbol{R}$ is a rotation/reflection/permutation matrix. The direction of the eigenvectors of the covariance matrix is changed, but the eigenvalues stay the same.

c) 70%. This is just combination of (a) and (b). All data points are scaled by 5 (i.e. eigenvalues of $\boldsymbol{X}^T\boldsymbol{X}$ are all scaled by 25), and some dimensions are reflected around origin, but the fraction of variance explained by the first K components stays the same.

d) We cannot tell without additional information. since each column (i.e. each dimension) is scaled by a different amount.

e) 70%. All data points are shifted by $\boldsymbol{\mu}$. But since we center the data as the first step of PCA, shifting has no effect.

f) 100%. Since $\text{rank}(\boldsymbol{A}) = 5$, $\text{rank}(\boldsymbol{Y}_6) \leq 5$ as well. This means that the data lies in a ≤ 5 dimensional subspace, and the first 5 principal components capture all the variance.

11 Gaussian Mixture Models

Problem 11 [(2+5+2)=9 points] Consider two random variables $\boldsymbol{x} \in \mathbb{R}^D$ and $\boldsymbol{y} \in \mathbb{R}^D$ distributed according to two different Gaussian mixture models

$$p(\boldsymbol{x}|\boldsymbol{\theta}^X) = \sum_{k=1}^{K_X} \pi_k^X \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k^X, \boldsymbol{\Sigma}_k^X),$$
$$p(\boldsymbol{y}|\boldsymbol{\theta}^Y) = \sum_{l=1}^{K_Y} \pi_l^Y \mathcal{N}(\boldsymbol{y}|\boldsymbol{\mu}_l^Y, \boldsymbol{\Sigma}_l^Y),$$

The first mixture $p(\boldsymbol{x}|\boldsymbol{\theta}^X)$ consists of K_X components with parameters $\boldsymbol{\theta}_k^X = (\pi_k^X, \boldsymbol{\mu}_k^X, \boldsymbol{\Sigma}_k^X)$ for $k \in \{1, \ldots, K_X\}$. Similarly, $p(\boldsymbol{y}|\boldsymbol{\theta}^Y)$ consists of K_Y components with parameters $\boldsymbol{\theta}_l^Y = (\pi_l^Y, \boldsymbol{\mu}_l^Y, \boldsymbol{\Sigma}_l^Y)$ for $l \in \{1, \ldots, K_Y\}$.

We generate a new random variable $\boldsymbol{z} \in \mathbb{R}^D$ as $\boldsymbol{z} = \boldsymbol{x} + \boldsymbol{y}$.

a) Describe the generative process (process of drawing the samples) for $\boldsymbol{z}$.

- Draw k from the categorical distribution on $\{1, \ldots, K_X\}$ with probabilities from $\boldsymbol{\pi}^X$.
- Draw $\tilde{\boldsymbol{x}}$ from the normal distribution $N(\boldsymbol{\mu}_k^X, \boldsymbol{\Sigma}_k^X)$
- Draw l from the categorical distribution on $\{1, \ldots, K_Y\}$ with probabilities from $\boldsymbol{\pi}^Y$.
- Draw $\tilde{\boldsymbol{y}}$ from the normal distribution $N(\boldsymbol{\mu}_l^Y, \boldsymbol{\Sigma}_l^Y)$
- Return $\tilde{\boldsymbol{z}} := \tilde{\boldsymbol{x}} + \tilde{\boldsymbol{y}}$ as a sample from $\boldsymbol{z}$.

(Lecture 11, slide 18)

b) Explain in a few sentences why $p(\boldsymbol{z}|\boldsymbol{\theta}^{\boldsymbol{X}}, \boldsymbol{\theta}^{\boldsymbol{Y}})$ is again a mixture of Gaussians.

Let $\boldsymbol{x}$ be drawn from the component k of $p(\boldsymbol{x} \mid \boldsymbol{\theta}^X)$ and $\boldsymbol{y}$ be drawn from the component l of $p(\boldsymbol{y} \mid \boldsymbol{\theta}^Y)$. Then $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_k^X, \boldsymbol{\Sigma}_k^X)$ and $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}_l^Y, \boldsymbol{\Sigma}_l^Y)$. Since $\boldsymbol{z}$ is a sum of two normally distributed random variables, it also follows normal distribution $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}_k^X + \boldsymbol{\mu}_l^Y, \boldsymbol{\Sigma}_k^X + \boldsymbol{\Sigma}_l^Y)$. There are $K_X \cdot K_Y$ such possible (k, l) combinations, each having probability $\pi_k^X \pi_l^Y$ respectively.

That is, $p(\boldsymbol{z} \mid \boldsymbol{\theta}^X, \boldsymbol{\theta}^Y)$ is a mixture of $K_X K_Y$ gaussians.

c) Write down the probability density function $p(\boldsymbol{z}|\boldsymbol{\theta}^X, \boldsymbol{\theta}^Y)$ of $\boldsymbol{z}$.
It's enough to just state the answer (no need to show the derivation).

$$p(\boldsymbol{z}|\boldsymbol{\theta}^X, \boldsymbol{\theta}^Y) = \sum_{k=1}^{K_X} \sum_{l=1}^{K_Y} \pi_k^X \pi_l^Y \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_k^X + \boldsymbol{\mu}_l^Y, \boldsymbol{\Sigma}_k^X + \boldsymbol{\Sigma}_l^Y).$$

12 Variational Inference

The exponential distribution with a scale parameter $\alpha > 0$ is defined as

$$\text{Expo}(\theta \mid \alpha) = \begin{cases} \frac{1}{\alpha}\exp(-\frac{1}{\alpha}\theta) & \text{if } \theta \geq 0 \\ 0 & \text{else} \end{cases}, \qquad \mathbb{E}[x] = \alpha, \qquad \mathbb{E}\left[x^2\right] = 2\alpha^2$$

Problem 12 [(4+2+2)=8 points] Consider the following probabilistic model.

$$p(z) = \text{Expo}(z \mid 1)$$
$$p(x \mid z) = \mathcal{N}(x \mid z, 1)$$

We want to approximate the posterior distribution $p(z \mid x)$ using a variational distribution

$$q(z \mid \beta) = \text{Expo}(z \mid \beta)$$

a) Write down a closed-form for ELBO $\mathcal{L}(\beta)$ and simplify it as far as you can. You can ignore the terms that are constant in β.

Plugging in the definition of ELBO

$$\begin{aligned}
\mathcal{L}(\beta) &= \mathbb{E}_q[\log p(x,z) - \log q(z \mid \beta)] \\
&= \mathbb{E}_q[\log p(x \mid z) + \log p(z) - \log q(z \mid \beta)] \\
&= \mathbb{E}_q\left[-\frac{1}{2}(x-z)^2 - z + \frac{1}{\beta}z\right] + \text{const.} \\
&= \mathbb{E}_q\left[xz - \frac{1}{2}z^2 - z - \log\frac{1}{\beta} + \frac{1}{\beta}z\right] + \text{const.} \\
&= x\mathbb{E}_q[z] - \frac{1}{2}\mathbb{E}_q\left[z^2\right] - \mathbb{E}_q[z] + \log\beta + \frac{1}{\beta}\mathbb{E}_q[z] + \text{const.} \\
&= x\beta - \beta^2 - \beta + \log\beta + 1 + \text{const.} \\
&= -\beta^2 + (1-x)\beta + \log\beta + \text{const.}
\end{aligned}$$

(HW12, Problem 2)

b) Is the ELBO convex in β? Justify your answer.

$-\beta^2$, $(1-x)\beta$ and $\log\beta$ are all concave functions of β, so their sum is also concave. Hence $\mathcal{L}(\beta)$ is not convex in β.

(Lecture 6, slide 17)

c) Outline the main steps for solving the optimization problem

$$\min_{\beta} \ \mathbb{KL}(q(z \mid \beta) \parallel p(z \mid x))$$

You don't need to perform the actual computations, just clearly describe each step.

Does this optimization problem have a closed-form solution? Why or why not?

Minimizing KL divergence

$$\min_{\beta} \ \mathbb{KL}(q(z \mid \beta) \parallel p(z \mid x))$$

is equivalent to maximizing the ELBO

$$\max_{\beta} \ \mathbb{E}_q \left[\log p(x, z) - \log q(z \mid \beta)\right]$$

(Lecture 12, slide 15)

We already showed that ELBO is concave in β for this model, so we simply need to compute the gradient w.r.t. β and set it to zero

$$\nabla_\beta \mathcal{L}(\beta) = -2\beta + (1 - x) + \frac{1}{\beta} \stackrel{!}{=} 0$$

$$-2\beta^2 + (1 - x)\beta + 1 \stackrel{!}{=} 0$$

Solving the quadratic equation and choosing the positive root yields the optimal β.

(HW12, Problem 2)