Machine Learning I Exam from 24.09.2020

This document is not official; it is not endorsed by the university or the lecturer.

120 minutes, no auxiliary tools allowed, $20 + 15 + 25 + 20 + 20 = 100$ points.

1. Multiple choice ($4 \times 5 = 20$ Points)

Answer the following multiple choice questions.

(a) The Bayes error is

- ☐ the lowest error of a linear classifier.
- ☐ the expected error of a random linear classifier.
- ■ the error of any nonlinear classifier.
- ☐ the error of a naive Bayes classifier .

(b) The Fisher linear discriminant find the projection $y = w^\top x$ of the data that maximises

- ☐ the margin between the two data generating distributions.
- ☐ the within-class variance divided by the between-class variance.
- ☐ the margin between the means of the data generating distributions.
- ■ the between-class variance divided by the within-class variance.

(c) A biased estimator is used to

- ■ make the estimator less affected by the sampling of the data.
- ☐ make the estimation procedure more sensitive to the sample data.
- ☐ reduce the risk of underfitting the data.
- ☐ None of the above, an unbiased estimator is always better.

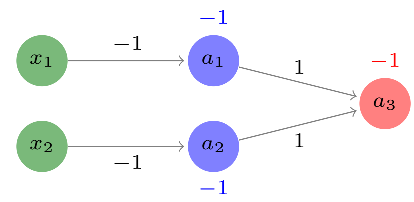
(d) Let $x_1, \ldots, x_N \in \mathbb{R}^d$ be unlabelled observations. Consider a Gaussian kernel and its Gram matrix $K \in \mathbb{R}^{N \times N}$. Which is always true?

- ☐ $K^T K = I$.
- ☐ $KK^T = I$.
- ■ $\forall u \in \mathbb{R}^N \; uKu \geq 0$.
- ☐ $\forall u \in \mathbb{R}^N \; uKu \leq 0$.

2. Neural Networks ($10 + 5 = 15$ Points)

(a) Build a neural network that models the function $f\colon \mathbb{R}^2 \to \{0,1\}$, $x \mapsto \mathbb{1}_{\min(x_1,x_2)\leq -1}(x)$ with at most three neurons of the form $a_j = \text{step}\left(\sum_i w_{ij}a_i + b_j\right)$, where $\text{step}(z) := \mathbb{1}_{\{z\geq 0\}}(z)$. State weights and biases.

Define $a_1 = \text{step}(-x_1 - 1)$ and $a_2 = \text{step}(-x_2 - 1)$ to check if $x_1 \leq -1$ and $x_2 \leq -1$. If (at least) one of them gives 1, we want the output to be one and zero else. Thus $a_3 = \text{step}(a_1 + a_2 - 1)$.



(b) State the number of neurons needed to build a neural network that models $f\colon \mathbb{R}^d \to \{0,1\}$, $x \mapsto \mathbb{1}_{\|x\|_\infty \leq 5}(x)$ and describe the weights and bias of one such neurons.

We need $2d+1$ neurons. We have that $\|x\|_\infty \leq 5$ if and only if $-5 \leq x_k \leq 5$ for all $k \in \{1,\dots,d\}$. For $k \in \{1,\dots,d\}$, we thus take $a_{2k-1} = \text{step}(5 - x_k)$ (to check that $x_k \leq 5$) and $a_{2k} = \text{step}(x_k + 5)$ (to check that $x_k \geq -5$). The output neuron is $a_{2d+1} = \text{step}(\sum_{k=1}^{2d} \frac{1}{2d} a_k - 1)$, as we only want to output 1 if all other a_j give 1.

3. Maximum likelihood and Bayes ($5 \times 5 = 25$ Points)

People queue at the post office and their i.i.d processing times are $D = (x_1, x_2, x_3) = (1, 1, 2)$. The data generating distribution is $P(x_i = k) = (1-\theta)^{k-1}\theta$, where $k \in \mathbb{N} \cup \{\infty\}$ and $\theta \in [0,1]$ is unknown.

(a) State likelihood function $P(D|\theta)$.

$P(D|\theta) = (1-\theta)^{1-1}\theta \cdot (1-\theta)^{1-1}\theta \cdot (1-\theta)^{2-1}\theta = \theta^3(1-\theta)$.

(b) Find the maximum likelihood parameter $\hat{\theta}$.

We have $\hat{\theta} = \arg\max_\theta P(D|\theta)$. We have $\frac{\mathrm{d}}{\mathrm{d}\theta}\theta^3(1-\theta) = 3\theta^2 - 4\theta^3$, so $\theta = 0$ or $\theta = \frac{3}{4}$. We also have to check the boundary of the definition domain of $P(D|\theta)$: we have $P(D|0) = 0 = P(D|1) < P(D|\frac{3}{4}) = \frac{27}{64}$, so $\hat{\theta} = \frac{3}{4}$.

(c) Evaluate $P(x_4 > 1|\hat{\theta})$.

Since x_4 can be every integer between 2 and ∞, we have

$$
\begin{aligned}
P(x_4 > 1|\hat{\theta}) &= \sum_{k=1}^{\infty} P(x_i = k) = \sum_{k=1}^{\infty} \left(1 - \hat{\theta}\right)^{k-1} \hat{\theta} = \sum_{k=2}^{\infty} \left(1 - \frac{3}{4}\right)^{k-1} \frac{3}{4} \\
&= \frac{3}{4} \sum_{k=1}^{\infty} \left(\frac{1}{4}\right)^k = \frac{3}{4} \cdot \frac{1}{3} = \frac{1}{4}.
\end{aligned}
$$

The sum is a geometric series so we can get the finite expression $\frac{1}{3}$ for it.

Simpler computation using the complement:

$$P(x_4 > 1|\hat{\theta}) = 1 - P(x_4 = 1|\hat{\theta}) = 1 - (1-\hat{\theta})^{1-1}\hat{\theta} = 1 - \hat{\theta} = 1 - \frac{3}{4} = \frac{1}{4}.$$

We now adopt a Bayesian view point on this problem, where we assume a prior distribution for the parameter θ to be defined as:

$$p(\theta) = \begin{cases} 1, & \theta \in [0,1], \\ 0 & \text{else.} \end{cases}$$

(d) Show that the posterior distribution $p(\theta|D)$ is $20(1-\theta)\theta^3$ for $\theta \in [0,1]$ and zero elsewhere.

By the theorem of Bayes and the law of total probability we have

$$\begin{aligned} p(\theta|D) &= \frac{p(D|\theta)p(\theta)}{p(D)} = \frac{p(D|\theta)p(\theta)}{\int_{\mathbb{R}} p(D|\theta)p(\theta)\,\mathrm{d}\theta} = \frac{\theta^3(1-\theta)\cdot \mathbb{1}_{[0,1]}(\theta)}{\int_0^1 \theta^3(1-\theta)\,\mathrm{d}\theta} \\ &= \frac{\theta^3(1-\theta)\cdot \mathbb{1}_{[0,1]}(\theta)}{\frac{1}{20}} = 20(1-\theta)\theta^3 \cdot \mathbb{1}_{[0,1]}(\theta) \end{aligned}$$

(e) Evaluate $P(x_4 > 1|D) = \int p(x|\theta)p(\theta|D)\,\mathrm{d}\theta$.

We have

$$\begin{aligned} P(x_4 > 1|D) &= 1 - P(x_4 = 1|\hat{\theta}) = 1 - \int 20(1-\theta)\theta^3 \cdot \mathbb{1}_{[0,1]}(\theta)\cdot\theta(1-\theta)^{1-1}\,\mathrm{d}\theta \\ &= 1 - 20\int_0^1 \theta^4(1-\theta)\,\mathrm{d}\theta = 1 - 20\int_0^1 \theta^4 - \theta^5\,\mathrm{d}\theta = 1 - 20\left(\frac{1}{5} - \frac{1}{6}\right) \\ &= 1 - \frac{2}{3} = \frac{1}{3} \end{aligned}$$

4. Lagrange multipliers ($4 \times 5 = 20$ Points)

Let $\Sigma \in \mathbb{R}^{d\times d}$ be a positive semidefinite matrix. Consider the constrained maximisation problem:

$$\max_{w\in\mathbb{R}^d} \|w\|^2 \quad \text{subject to} \quad w^\top \Sigma^{-1} w = 1$$

(a) State the Lagrange function.

$L(w,\lambda) := \|w\|^2 + \lambda(1 - w^\top\Sigma^{-1}w).$

(b) Show that the problem is an eigenvalue problem of Σ.

For w to be optimal, we need

$$\frac{\partial L(w,\lambda)}{\partial w} = 2w - 2\lambda\Sigma^{-1}w \overset{!}{=} 0 \iff w = \lambda\Sigma^{-1}w \iff \Sigma w = \lambda w,$$

so w has to be a eigenvector of Σ with eigenvalue λ.

(c) Show that the solution is the eigenvector associated to the highest eigenvalue of Σ.

From the constraint $w^\mathsf{T}\Sigma^{-1}w = 1$ and $w = \lambda\Sigma^{-1}w$ we get (as Σ is symmetric)

$$\|w\|^2 = w^\mathsf{T}w = \lambda w^\mathsf{T}\Sigma^{-1}w = \lambda.$$

Thus the value of the eigenvalue coincides with the quantity we want to maximise.

(d) Let $w_1, \dots, w_T$ be a sequence of vectors where w_t is obtained from w_{t-1} as the solution of the constraint problem

$$\max_{z\in\mathbb{R}^d} z^\mathsf{T}w_{t-1} \quad \text{subject to} \quad z^\mathsf{T}\Sigma^{-1}z = 1.$$

Find a closed form solution of w_t as a function of w_{t-1}.

The Lagrangian is

$$L(z,\lambda) \coloneqq z^\mathsf{T}w_{t-1} + \lambda(1 - z^\mathsf{T}\Sigma^{-1}z).$$

In order for z to be optimal, we require

$$\frac{\partial L(z,\lambda)}{\partial z} = w_{t-1} - 2\lambda\Sigma^{-1}z \overset{!}{=} 0 \iff w_{t-1} = 2\lambda\Sigma^{-1}z \iff z = \frac{1}{2\lambda}\Sigma w_{t-1}.$$

Plugging the second last equality into the constraint $z^\mathsf{T}\Sigma^{-1}z = 1$, we get

$$z^\mathsf{T}w_{t-1} = 2\lambda z^\mathsf{T}\Sigma^{-1}z = 2\lambda$$

and using the last equality we get

$$2\lambda = z^\mathsf{T}w_{t-1} = \frac{1}{2\lambda}w_{t-1}^\mathsf{T}\Sigma w_{t-1},$$

implying

$$2\lambda = \sqrt{w_{t-1}^\mathsf{T}\Sigma w_{t-1}},$$

as Σ is positive semidefinite (so we don't have to consider $-\sqrt{\dots}$). We thus get

$$w_t = z = \frac{\Sigma w_{t-1}}{\sqrt{w_{t-1}^\mathsf{T}\Sigma w_{t-1}}} = \frac{\Sigma w_{t-1}}{\|\Sigma w_{t-1}\|_{\Sigma^{-1}}} \quad \text{with } \|x\|_{\Sigma^{-1}}^2 \coloneqq x^\mathsf{T}\Sigma^{-1}x.$$

5. Ridge regression $(10 + 10 = 20$ Points)

Consider the problem

$$\min_{w \in \mathbb{R}^d} \|y - Xw\|^2 \quad \text{subject to} \quad \|w\|_\infty \leq C,$$

where $C > 0$ is a constant, $y \in \mathbb{R}^N$ and $X \in \mathbb{R}^{N \times d}$ is the data matrix.

(a) Show that the problem is equivalent to

$$\min_{w \in \mathbb{R}^d} w^\top X^\top X w - 2y^\top X w \quad \text{subject to} \quad -C \leq w_i \leq C \ \forall i \in \{1, \ldots, d\}$$

We have

$$\begin{aligned} \|y - Xw\|^2 = (y - Xw)^\top (y - Xw) &= y^\top y - y^\top Xw - (Xw)^\top y + (Xw)^\top Xw \\ &= y^\top y - 2y^\top Xw + w^\top X^\top X w. \end{aligned}$$

Since $y^\top y$ is independent of w, we can neglect it when minimising over w. We have $y^\top Xw = (Xw)^\top y$, as it is a scalar and so it is equal to its transpose.

Furthermore, $\|w\|_\infty = \max\{|w_1|, \ldots, |w_d|\}$, so $\|w\|_\infty \leq C$ is equivalent to $|w_k| \leq C$ for all $k \in \{1, \ldots, d\}$, i.e. $-C \leq w_k \leq C$ for all $k \in \{1, \ldots, d\}$.

(b) At our disposal we have a quadratic solver `QP(Q, l, A, b)`, which solves the generic quadratic problem

$$\min_v v^\top Q v + \ell^\top v \quad \text{subject to} \quad Av \leq b.$$

Write the numpy code constructing the arrays Q, ℓ, A and b from X, y and C.

```
def Reg(X, y, C):
    Q = X.T.dot(X)
    l = - 2 * y.T.dot(X).T
    d = Q.shape[0]
    A = np.concatenate([np.identity(d), -1 * np.identity(d)], axis=0)
    b = C * np.ones(2 * d)
    t = QP(Q, l, A, b)
return t
```

The grey code was given.

Thanks to everyone contributing to this account of the exam and its solutions :)