Introduction to Machine Learning - Summer 2019
Final Exam

Instructions

1. There are 10 questions (each question is 10% of the total grade).
2. Provide full solutions (explain your answers).
3. You can keep the questions form with you (so don't write your solution on it).
4. You can use a draft notebook (you don't need to submit it).
5. Write your student ID on the notebook you are submitting.
6. Good Luck!

1 Estimation

Let $\hat{\theta}$ be an estimator of θ.

- Bias:
$$b\left(\hat{\theta}\right) = \mathbb{E}\left[\hat{\theta}\right] - \theta$$
- Variance:
$$\mathrm{Var}\left(\hat{\theta}\right) = \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]\right)^2\right]$$
- MSE:
$$\mathrm{MSE}\left(\hat{\theta}\right) \triangleq \mathbb{E}\left[\left(\hat{\theta} - \theta\right)^2\right]$$

Prove that:
$$\mathrm{MSE}\left(\hat{\theta}\right) = \mathrm{Var}\left(\hat{\theta}\right) + b^2\left(\hat{\theta}\right)$$

2 ML

The MLE $\hat{\theta}_{ML}$ of the parameter θ is defined by:
$$\hat{\theta}_{ML} \triangleq \arg\max_{\theta} \mathcal{L}\left(\theta\right) = \arg\max_{\theta} p\left(\{x_i\};\theta\right)$$
where $\mathcal{L}\left(\theta\right) = p\left(\{x_i\};\theta\right)$ is the likelihood function.

Consider the random variable X with the following probability density function:
$$f_X(x) = \begin{cases} \lambda x^{-2}\exp(-\lambda/x) & x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

A set $\mathcal{D} = \{x_i\}_{i=1}^N$ of i.i.d samples from f_X is given.

1. Write the log likelihood function $\ell(\lambda)$ and compute the maximum likelihood estimator $\hat{\lambda}_{ML}$ of λ.
2. Write the value of your estimation given two observations $x_1 = 2$, $x_2 = \frac{1}{2}$.

3 MAP

The MAP estimator $\hat{\theta}_{MAP}$ of the random variable θ is defined by:

$$\hat{\theta}_{MAP} \triangleq \arg\max_{\theta} p\left(\theta \middle| \{x_i\}\right) = \arg\max_{\theta} p\left(\{x_i\} \middle| \theta\right) p\left(\theta\right)$$

Consider a random variable $K \sim \text{Beta}(\alpha, \beta)$ (beta distribution) whose probability density function with parameters α, $\beta > 0$ is given by

$$f_K(k; \alpha\, \beta) = \begin{cases} C \cdot k^{\alpha-1}(1-k)^{\beta-1} & 0 \leq k \leq 1, \\ 0 & \text{else} \end{cases}$$

where C is some constant that depends on α, β.

1. Write an expression for C as a function of α, β. It may contain sums and integrals.

2. Consider the binary random variable X (a coin toss):

$$P_X\left(x\right) = \begin{cases} 1-k & x = 0, \qquad \text{(tails)} \\ k & x = 1, \qquad \text{(heads)} \end{cases}$$

   Write the probability to get one *heads* and one *tails* out of two independent coin flips (as a function of k).

3. Assume we got one *heads* and one *tails* out of two coin flips and that $k \sim \text{Beta}(3, 1)$. What is the MAP estimator $\hat{k}_{MAP}$ of k?

4 Non-parametric estimation

The CDF of the random variable X is given by:

$$F_X\left(x\right) = \Pr\left\{X \leq x\right\}$$

Given $\{x_i\}_{i=1}^{N}$, N i.i.d realizations of X we define the following estimator for F_X:

$$\hat{F}_X\left(x_0\right) = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{I}\left\{x_i \leq x_0\right\}$$

1. Compute the bias of $\hat{F}_X\left(x_0\right)$ (for a fixed x_0).

2. Compute the MSE $\left(\hat{F}_X\left(x_0\right)\right)$ (for a fixed x_0).

3. Find the limit value of the MSE as $N \to \infty$:

$$\text{MSE}\left(\hat{F}_X\left(x_0\right)\right) \underset{N\to\infty}{\longrightarrow} ?$$

5 PCA I

Let $\{\boldsymbol{x}_i\}_{i=1}^N$ be a set of vectors such that $\boldsymbol{x}_i \in \mathbb{R}^d$.

- Empirical mean and covariance:

$$\boldsymbol{\mu}_x = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i \triangleq \overline{\boldsymbol{x}}, \qquad \boldsymbol{\Sigma}_x = \frac{1}{N}\sum_{i=1}^{N} \left(\boldsymbol{x}_i - \boldsymbol{\mu}_x\right)\left(\boldsymbol{x}_i - \boldsymbol{\mu}_x\right)^T$$

- The eigen decomposition of $\boldsymbol{\Sigma}_x$ is given by:

$$\boldsymbol{\Sigma}_x = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$$

  where $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$, and $\boldsymbol{\Lambda}$ is a diagonal matrix with non-negative elements.

We define the following map:

$$\phi\left(\boldsymbol{x}\right) = \boldsymbol{U}^T\left(\boldsymbol{x} - \boldsymbol{\mu}_x\right)$$

and denote:

$$\boldsymbol{y}_i = \phi\left(\boldsymbol{x}_i\right) = \boldsymbol{U}^T\left(\boldsymbol{x}_i - \boldsymbol{\mu}_x\right), \qquad \forall i \in \{1, 2, \ldots, N\}$$

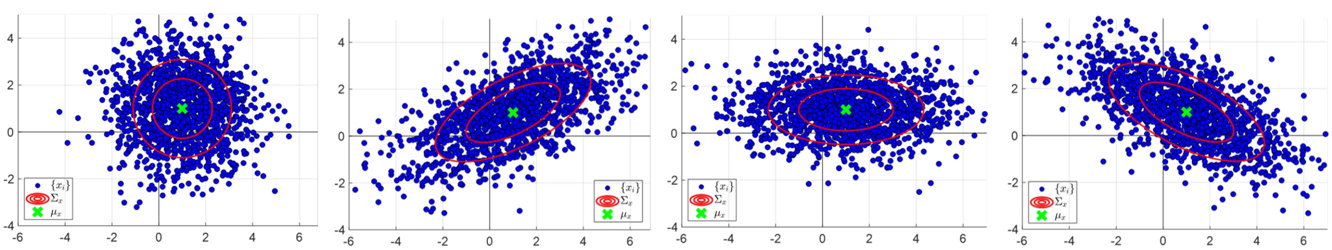
Prove that:

1. $\boldsymbol{\mu}_y = \boldsymbol{0}$.
2. $\boldsymbol{\Sigma}_y$ is a diagonal matrix.
3. $\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_2 = \left\|\boldsymbol{y}_i - \boldsymbol{y}_j\right\|_2$ for all i, j.

6 PCA II + K-means

6.1

Match between each of the four data sets and their corresponding covariance matrix:



(a) (b) (c) (d)

$$(1)\ \boldsymbol{\Sigma}_x = \begin{bmatrix} 5 & 0 \\ 0 & 2 \end{bmatrix}, \qquad (2)\ \boldsymbol{\Sigma}_x = \begin{bmatrix} 5 & 2 \\ 2 & 2 \end{bmatrix}, \qquad (3)\ \boldsymbol{\Sigma}_x = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, \qquad (4)\boldsymbol{\Sigma}_x = \begin{bmatrix} 5 & -2 \\ -2 & 2 \end{bmatrix}$$

6.2

In K-means, we seek to minimize the following objective function:

$$\min_{\{\mathcal{C}_k\},\{\boldsymbol{\mu}_k\}} \sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in \mathcal{C}_k} \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|_2^2$$

Where $K \in \mathbb{N}$ is the desired number of clusters.

Consider the set $\mathcal{D} = \left\{\boldsymbol{x}_i \in \mathbb{R}^d\right\}_{i=1}^{N}$.
Given fixed and known clusters $\mathcal{C}_k$ such that $\bigsqcup_{k=1}^{K} \mathcal{C}_k = \mathcal{D}$,
find the optimal centroids $\left\{\boldsymbol{\mu}_k \in \mathbb{R}^d\right\}$ which minimize the objective function.

7 MAP Classifier

- The MAP classifier is given by:

$$f_{\text{MAP}}(\boldsymbol{x}) = \arg\max_{C_k \in \mathcal{Y}} p(\boldsymbol{x}|C_k)\, P_{\mathcal{Y}}(C_k)$$

In the binary case ($\mathcal{Y} = \{C_1, C_2\}$), the decision rule is given by:

$$p(\boldsymbol{x}|C_1)\, P_{\mathcal{Y}}(C_1) \underset{C_2}{\overset{C_1}{\gtrless}} p(\boldsymbol{x}|C_2)\, P_{\mathcal{Y}}(C_2)$$

and the decision boundary is given by:

$$p(\boldsymbol{x}|C_1)\, P_{\mathcal{Y}}(C_1) = p(\boldsymbol{x}|C_2)\, P_{\mathcal{Y}}(C_2)$$

- The multivariate Gaussian distribution is given by ($X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$):

$$P_X(\boldsymbol{x}) = |2\pi\boldsymbol{\Sigma}|^{-1} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right), \qquad \boldsymbol{\mu} \in \mathbb{R}^d,\ \boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$$

Consider the random vector $X \in \mathbb{R}^d$ such that:

$$X|C_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{I})$$
$$X|C_2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{I})$$

with the prior:

$$P_{\mathcal{Y}}(C_1) = p_1, \qquad P_{\mathcal{Y}}(C_2) = 1 - p_1$$

Show that the decision boundary is linear, that is, it can be written as

$$\boldsymbol{w}^T \boldsymbol{x} - b = 0$$

Write $\boldsymbol{w}$ and b explicitly as a function of $p_1, \boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$.

8 Regression

Given a set of points $\{x_i, y_i\}_{i=1}^{N}$, the regression L_2 error of the function $\hat{f}$ is given by:

$$L_2\left(\hat{f}\right) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{f}\left(x_i\right)\right)^2$$

The M order polynomial function is given by:

$$\hat{f}\left(x\right) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{m=0}^{M} w_m x^m$$

where $\{w_m\}_{m=0}^{M}$ are the polynomial coefficients.
For a given set of points $\{x_i, y_i\}_{i=1}^{N}$, find the optimal $\{w_m\}_{m=0}^{M}$ which minimize the L_2 error.
The solution can be written in a vector form
(make sure to explain the dimensions and content of each matrix or vector).

Hint: define $\phi\left(x\right) = \begin{bmatrix}1 & x & x^2 & \cdots & x^M\end{bmatrix}$.

9 Linear SVM

Given a training set $\{\boldsymbol{x}_i, y_i\}_{i=1}^{N}$, the support vector machine (SVM) classification problem is given by:

$$\min_{\boldsymbol{w},\boldsymbol{b}} \frac{1}{2}||\boldsymbol{w}||_2^2$$
$$\text{s.t. } y_i\left(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b\right) \geq 1. \qquad i = 1, 2, ..., N.$$

Consider a training set $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^{N}$ where $y_i \in \{-1, 1\}$ and $\mathcal{D}$ is linearly separable.
Denote the optimal SVM solution by $\boldsymbol{w}^*$ and b^*.
We define a new training set $\tilde{\mathcal{D}} = \{\tilde{x}_i, y_i\}_{i=1}^{N}$ using the following transformation:

$$\tilde{x}_i = \langle \boldsymbol{w}^*, \boldsymbol{x}_i\rangle + b^*, \quad i = 1, 2, ..., N.$$

Is $\tilde{\mathcal{D}}$ linearly separable?
If so, write an expression for the **optimal** SVM solution $(\tilde{w}, \tilde{b})$ which perfectly separates $\tilde{D}$.
If $\tilde{D}$ is linearly non-separable, provide an example.

10 Kernels

- A **kernel** function $k : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}$ satisfies:
  1. Symmetry: $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = k(\boldsymbol{x}_j, \boldsymbol{x}_i)$
  2. k is positive definite, that is, the matrix $\boldsymbol{K}[i,j] = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is PSD for any set $\{\boldsymbol{x}_i\}_i$
- A kernel function k can be written as (for some $M \in \{\mathbb{N} \cup \infty\}$):

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{m=1}^{M} \phi_m(\boldsymbol{x}_i)\,\phi_m(\boldsymbol{x}_j) = \langle \boldsymbol{\phi}(\boldsymbol{x}_i), \boldsymbol{\phi}(\boldsymbol{x}_j) \rangle$$

- Given a training set $\{\boldsymbol{x}_i, y_i\}_{i=1}^{N}$, the optimal weights for the SVM task are given by:

$$\boldsymbol{w}^* = \sum_{i=1}^{N} \alpha_i^* y_i \boldsymbol{x}_i$$

  where $\{\alpha_i^*\}_i$ are the solution of the dual problem.
  Hence, $\langle \boldsymbol{w}^*, \boldsymbol{x}_0 \rangle = \sum_{i=1}^{N} \alpha_i^* y_i \langle \boldsymbol{x}_i, \boldsymbol{x}_0 \rangle$ depends only on the inner product between the training vectors.
  This allows to extend SVM to the non-linear case in which:

$$\langle \boldsymbol{w}^*, \boldsymbol{x}_0 \rangle = \sum_{i=1}^{N} \alpha_i^* y_i k(\boldsymbol{x}_i, \boldsymbol{x}_0)$$

10.1

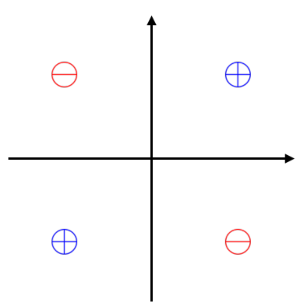
Show that the following kernel can be written as an inner product, namely:

$$k(x,y) = (1 + x \cdot y)^2 = \langle \phi(x), \phi(y) \rangle, \qquad x, y \in \mathbb{R}$$

Write the transformation $\phi(x)$ explicitly.

10.2

Consider the following training set:



For each kernel, write if the classification task is linear separable or linear non-separable.

1. $k(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^T \boldsymbol{y}$
2. $k(\boldsymbol{x}, \boldsymbol{y}) = \left(\boldsymbol{x}^T \boldsymbol{y}\right)^2$
3. $k(\boldsymbol{x}, \boldsymbol{y}) = \left(1 + \boldsymbol{x}^T \boldsymbol{y}\right)^2$
4. $k(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|_2^2}{2\sigma^2}\right)$
   (you may chose the value of σ)