Introduction to Machine Learning - Summer 2018
Final Exam

Instructions

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

1 (10%) ML

The MLE $\hat{\theta}_{ML}$ of the parameter θ is defined by:

$$\hat{\theta}_{ML} \triangleq \arg\max_{\theta} p\left(\{x_i\};\theta\right)$$

Consider a Gaussian random variable $Y \sim \mathcal{N}(\mu, \sigma^2)$ where μ and σ^2 are unknowns.
We define a new random variable as $X = e^Y$.
The probability density function of X is given by:

$$p_X\left(x;\mu,\sigma^2\right) = \begin{cases} \frac{1}{x}\cdot\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\left(\log(x)-\mu\right)^2}{2\sigma^2}\right) & x>0 \\ 0 & x\leq 0 \end{cases}$$

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

- Write the log-likelihood function $\ell\left(\mu,\sigma^2\right)$ and compute the maximum likelihood estimator for $\exp(\mu)$.
- Write the value of your estimation given two observations $\{x_1 = 3, x_2 = 2\}$.

2 (10%) MAP

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

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

Reminder

- $X \sim \text{Poisson}(\lambda) \quad \Rightarrow \quad P(X=k) = \frac{\lambda^k}{k!}e^{-\lambda}$ where $\lambda > 0$ and $k = 0,1,2,...$
- $Y \sim \text{Gamma}(\alpha,\beta) \quad \Rightarrow \quad p_Y(y) = \frac{\beta^\alpha}{\Gamma(\alpha)}y^{\alpha-1}e^{-\beta y}$ where $y,\alpha,\beta > 0$.

It is known that $\lambda \sim \text{Gamma}(\alpha,\beta)$ and given λ we have: $X|\lambda \sim \text{Possion}(\lambda)$.
Consider a set of i.i.d samples $\mathcal{D} = \{x_i\}_{i=1}^N$ drawn from X.

- Find the MAP estimator $\hat{\lambda}_{MAP}$ for λ and write it as a function of α,β and $\{x_i\}$.
- Write an expression for $\hat{\lambda}_{MAP}$ as $N \to \infty$.

3 (10%) Bayes Classifier

Remainder

Distributions Table

Distribution	Notation	Support	PDF	Mean	Variance
Uniform	$x \sim U[a,b]$	$x \in [a,b]$	$f(x) = \frac{1}{b-a}$	$\frac{b+a}{2}$	$\frac{1}{12}(b-a)^2$
Normal	$x \sim N(\mu,\sigma^2)$	$\Re$	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
Exponential	$x \sim \text{Exp}(\lambda)$	$0 \le x \in \Re$	$f(x) = \frac{1}{\lambda} e^{-x/\lambda}$	λ	λ^2

A decision problem is given, in which the input space is the real numbers, $X = \mathbb{R}$.
Input examples belong to one of three classes: $\Omega = \{e, g, u\}$. The conditional distributions for the three classes are:

- Class u: $x \sim U[0,3]$.
- Class g: $x \sim \mathcal{N}(2,4)$.
- Class e: $x \sim \text{Exp}(2)$.

The prior distributions are given by

- $p(u) = 0.3$.
- $p(g) = 0.3$.
- $p(e) = 0.4$.

What is the optimal Bayes classifier of the state, given a single input x?

- Write a function from the real numbers to Ω:

$$\hat{\omega}(x) = \begin{cases} A & x < a \\ B & a \le x < b \\ C & b \le x < c \\ D & c \le x < d \\ E & x \ge d \end{cases}$$

  Determine $a, b, c, d \in \mathbb{R}$ and $A, B, C, D, E \in \Omega$.

Hint: Sketch the conditional distributions and use the following facts:

- $\frac{1}{\sqrt{8\pi}} \approx 0.1995$.
- $2\ln(2) \approx 1.38$.
- $x^2 - 8x + 4 - 8\Big(\ln(1.5) - \frac{1}{2}\ln(8\pi)\Big) \ge 0 \quad \Rightarrow \quad x \le 2.468$ or $x \ge 5.532$.

4 (10%) Histogram

Given N i.i.d realizations of the random variable $X \in \mathcal{X}$: $\mathcal{D} = \{x_i\}_{i=1}^N$.
We can split the domain $\mathcal{X}$ to K disjoint intervals:

$$\mathcal{X} = \dot{\bigcup_k} R_k$$

The histogram estimation at the point $x_0 \in R_k$ is given by:

$$\hat{p}_X(x_0) = \frac{1}{N} \cdot \frac{1}{|R_k|} \sum_{i=1}^{N} \boldsymbol{I}\{x_i \in R_k\}$$

where $|R_k|$ is the length of the interval R_k.

For $x_0 \in R_k = (a, b]$, compute:

$$\mathbb{E}[\hat{p}_X(x_0)] = ?$$

Write your expression as a function of the true CDF F_X, and the parameters N, a, b and $\{x_i\}$.

5 (10%) PCA I

Consider the set of points $\mathcal{D} = \left\{\boldsymbol{x}_i \in \mathbb{R}^D\right\}_{i=1}^N$
The empirical mean is given by:

$$\boldsymbol{\mu}_x = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x}_i$$

The empirical covariance is given by:

$$\boldsymbol{\Sigma}_x = \frac{1}{N} \sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu}_x)(\boldsymbol{x}_i - \boldsymbol{\mu}_x)^T$$

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

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

where $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{I}$ and $\boldsymbol{\Lambda}$ is a diagonal matrix.
The PCA transformation is given by:

$$\boldsymbol{y}_i = \boldsymbol{U}^T(\boldsymbol{x}_i - \boldsymbol{\mu}_x)$$

- Prove that for all $i, j \in \{1, 2, \ldots, N\}$:
$$\left\|\boldsymbol{y}_i - \boldsymbol{y}_j\right\|_2 = \left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_2$$
- Prove that $\boldsymbol{\Sigma}_y$, the covariance of $\{\boldsymbol{y}_i\}$ is a diagonal matrix.

6 (10%) PCA II

Consider the set of points $\left\{\boldsymbol{x}_i \in \mathbb{R}^D\right\}_{i=1}^N$.
It is known that the empirical mean is $\boldsymbol{\mu}_x = 0$ and their empirical convarince matrix is given by its eigen decomposition:

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

where $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{I}$ and $\boldsymbol{\Lambda}$ is a diagonal matrix.
Given the matrix $\boldsymbol{U}_m \in \mathbb{R}^{D\times m}$ whose columns are the first $m < D$ principle components, we perform dimensionality reduction by applying to each sample the following transform:

$$\boldsymbol{y}_i = \boldsymbol{U}_m^T \boldsymbol{x}_i \in \mathbb{R}^m.$$

In this question, we create a new set of samples $\left\{\tilde{\boldsymbol{x}}_i \in \mathbb{R}^K\right\}_{i=1}^N$ for some $K > D$, by applying to each sample the following linear transformation:

$$\tilde{\boldsymbol{x}}_i = \boldsymbol{V}\boldsymbol{x}_i \in \mathbb{R}^K$$

where $\boldsymbol{V} \in \mathbb{R}^{K\times D}$ is a matrix which satisfies $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$ with $\boldsymbol{I} \in \mathbb{R}^{D\times D}$ being the identity matrix.
Applying PCA to $\left\{\tilde{\boldsymbol{x}}_i \in \mathbb{R}^K\right\}_{i=1}^N$ using the first m principle components provides a new set of representation vectors $\{\tilde{\boldsymbol{y}}_i \in \mathbb{R}^m\}_{i=1}^N$.
Write an expression for $\tilde{\boldsymbol{y}}_i$ as a function of $\boldsymbol{y}_i$ and give a detailed mathematical explanation.

7 (10%) K-Means

The objective function of K-means is given by:

$$J\left(\{\boldsymbol{\mu}_k\}, \{\mathcal{C}_k\}\right) = \sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in \mathcal{C}_k} \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|_2^2$$

where $\{\boldsymbol{\mu}_k\}_{k=1}^K$ are the clusters centroids and
the clusters $\{\mathcal{C}_k\}_{k=1}^K$ are disjoint subsets of the entire set $\{\boldsymbol{x}_i\}$.
The K-means algorithm is given by:

Algorithm 1 K-Means

1. Set initial centroids $\{\boldsymbol{\mu}_k\}$
2. Find the clusters $\{\mathcal{C}_k\}$ by:
$$\mathcal{C}_s = \left\{\boldsymbol{x}_i \;\middle|\; \|\boldsymbol{x}_i - \boldsymbol{\mu}_s\|_2^2 \le \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|_2^2\right\}$$
3. Update the centroids:
$$\boldsymbol{\mu}_s = \frac{1}{|\mathcal{C}_s|} \sum_{\boldsymbol{x}_i \in \mathcal{C}_s} \boldsymbol{x}_i$$
4. Repeat 2 - 3 until convergence.

Consider some iteration t.
We denote the value of the objective function **before** step 3 by: $J_0 = J\left(\{\boldsymbol{\mu}_k\}, \{\mathcal{C}_k\}\right)$,
where $\{\boldsymbol{\mu}_k\}$ are the centroids before the update step.
We denote the value of the objective function **after** step 3 by: $J_1 = J\left(\{\boldsymbol{m}_k\}, \{\mathcal{C}_k\}\right)$,
where $\{\boldsymbol{m}_k\}$ are the centroids after the update step.
Prove that:

$$J_1 \le J_0$$

Hint: show that $J_1 - J_0 \le 0$. It is also enough to consider a single cluster $k \in \{1, 2, \ldots, K\}$.

8 (10%) Perceptron Algorithm

The perceptron algorithm (without bias, namely: $b = 0$) is given by:

Algorithm 2 The Perceptron Algorithm

Input: Training set $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}$
Output: The linear classifiers parameters: $\tilde{\boldsymbol{w}}$.

1. Set $\boldsymbol{w}_1$ with some initial guess.
2. **for** $k = 1, 2, 3, \ldots$
   (a) Choose some $(\boldsymbol{x}_k, y_k) \in \mathcal{D}$
   (b) Compute:
   $$\hat{y}_k = \text{sign}\left(\boldsymbol{w}_k^T \boldsymbol{x}_k\right)$$
   (c) Update:
   $$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k + \frac{1}{2}\left(y_k - \hat{y}_k\right)\boldsymbol{x}_k$$

Consider a training set containing a single example: $\mathcal{D} = \{(\boldsymbol{x}_1, y_1)\}$ such that $\|\boldsymbol{x}_1\| = 1$ and $y_1 = 1$.
The algorithm is initialized with some $\boldsymbol{w}_1$ such that $\|\boldsymbol{w}_1\|_2 \leq 10$.

1. Find a tight lower bound for the number of iterations until convergence.
   Provide an example to prove the tightness.
2. Find a tight upper bound for the number of iterations until convergence.
   Provide an example to prove the tightness.

9 (10%) Regression

Given a training set:

$$\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^N$$

A linear model between the $\boldsymbol{x}_i$ and y_i is given by:

$$y_i = \boldsymbol{w}^T \boldsymbol{x}_i$$

Consider instead of the L_2 error, the weighted error which is given by:

$$L\left(\boldsymbol{w}\right) = \text{weighted-error} = \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\right)^T \boldsymbol{A}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\right)$$

where:

$$\boldsymbol{y} \triangleq \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}, \qquad \boldsymbol{X} = \begin{bmatrix} | & & | \\ \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_N \\ | & & | \end{bmatrix}^T$$

and $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ is the weight matrix.

1. Compute the gradient of $L\left(\boldsymbol{w}\right)$:
   $$\nabla_{\boldsymbol{w}} L = ?$$
2. Assuming some initial guess $\boldsymbol{w}_0$.
   Write the gradient descent update for a fixed step size μ:
   $$\boldsymbol{w}_1 = ?$$

10 (10%) Kernel function

Let $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ be a symmetric and positive definite matrix with the following eigen decomposition:

$$\boldsymbol{A} = \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^T$$

where $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{I}$ and $\boldsymbol{\Lambda}$ is a diagonal matrix with positive values (on the diagonal).
Consider the following kernel function:

$$k(\boldsymbol{x}, \boldsymbol{z}) = \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{z}$$

where $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^d$.
Find a transformation $\boldsymbol{\phi}(\boldsymbol{x})$ such that:

$$k(\boldsymbol{x}, \boldsymbol{z}) = \langle \boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{z}) \rangle$$