School of Electronics and Computer Science
University of Southampton

COMP6245 (2020/21): Foundations of Machine Learning **Exam**

Attempt all questions. Provide answers to all questions in a single page, neatly numbered in order. You may attach two further pages with any workings where useful, clearly numbered and in order of the questions. These additional pages will be looked at if the question requires any derivation and the answer you provided is incorrect.
Note the questions will have at least one correct answer. Where the question has more than one correct answer, you must select all the correct ones. For these questions, partial credit will usually not be given.
You should upload a maximum of three pages as a single pdf file.
Each question is worth *five marks*.

Date	TBD
Timing	Four hours within a 24 hour period

1. To solve a two-class classification problem with classes denotes $\mathcal{A}$ and $\mathcal{B}$, a classifier is designed to compute $y = \boldsymbol{x}^T A\, \boldsymbol{x} + \boldsymbol{b}^T \boldsymbol{x} + c$ on input features $\boldsymbol{x} \in \mathcal{R}^p$, with $A \in \mathcal{R}^{p \times p}$, $\boldsymbol{b} \in \mathcal{R}^p$ and $c \in \mathcal{R}$ as its parameters. Once the parameters are determined, a decision rule of the form $y > \theta \implies \mathcal{A}$ is applied. Which of the following conditions are necessary for the above classifier to be optimal.

1. The prior probabilities of the two classes are different: *i.e.* $p(\mathcal{A}) \neq p(\mathcal{B})$.
2. The class conditional densities are *mixture Gaussian distributions*: *i.e.* $p(\boldsymbol{x}|\mathcal{A} = \sum_{k=1}^{K_{\mathcal{A}}} \lambda_k \mathcal{N}(\boldsymbol{m}_k, \boldsymbol{\Sigma}_k)$ and likewise for class $\mathcal{B}$.
3. The class conditional probabilities are both Gaussian: *i.e.* $p(\boldsymbol{x}|\mathcal{A}) = \mathcal{N}(\boldsymbol{m}_{\mathcal{A}}, \boldsymbol{\Sigma}_{\mathcal{A}})$ and $p(\boldsymbol{x}|\mathcal{B}) = \mathcal{N}(\boldsymbol{m}_{\mathcal{B}}, \boldsymbol{\Sigma}_{\mathcal{B}})$, where $\boldsymbol{\Sigma}_{\mathcal{A}} \neq \boldsymbol{\Sigma}_{\mathcal{B}}$.
4. The class conditional probabilities are both Gaussian and should have identical an isotropic covariance metrices: *i.e.* $p(\boldsymbol{x}|\mathcal{A}) = \mathcal{N}(\boldsymbol{m}_{\mathcal{A}}, \boldsymbol{\Sigma}_{\mathcal{A}})$ and $p(\boldsymbol{x}|\mathcal{B}) = \mathcal{N}(\boldsymbol{m}_{\mathcal{B}}, \boldsymbol{\Sigma}_{\mathcal{B}})$, where $\boldsymbol{\Sigma}_{\mathcal{A}} = \boldsymbol{\Sigma}_{\mathcal{B}} = \sigma^2 \boldsymbol{I}$.

2. A variable of interest, x, has an exponential distribution:

$$p(x_i|\theta) = \begin{cases} \theta \exp(-\theta x_i) & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Given N samples, $x_1, x_2, ..., x_N$, drawn independently from $P(x|\theta)$, derive the maximum likelihood estimate of θ.

Your answer is:

1. $\widehat{\theta} = \frac{1}{\frac{1}{N}\sum_{n=1}^{N} x_n}$
2. $\widehat{\theta} = \prod_{n=1}^{N} x_n$
3. $\widehat{\theta} = \frac{N!}{2\pi} \sum_{n=1} N x_n$
4. $\widehat{\theta} = \frac{1}{\sqrt{2\pi}} \sum_{n=1}^{N} x_n$

3. You are tasked with predicting the market price of an asset using its past values and several variables relating to the underlying economy within which the business operates. The dataset given to you spans six months of daily trading (125 items) of 100 variables. You are required to split the data into a training set and an evaluation set of equal sizes and use a linear model as predictor. You attempt to solve the problem of estimating regression coefficients by

$$\boldsymbol{w} = \left(X^T X + \gamma \boldsymbol{I}\right)^{-1} X^T \boldsymbol{t}$$

Which of the following is/are true?

1. ✗ γ is referred to as momentum.
2. ✗ The $\gamma \boldsymbol{I}$ term in the formula is not necessary to solve this problem.
3. ✓ Inclusion of the $\gamma \boldsymbol{I}$ term is equivalent to *weight decay*, sometimes used in training neural networks.
4. ✓ I would advocate the use of a *Lasso* regularizer to solve the problem.

4. A researcher attempting to predict median house prices in different parts of Southampton from 15 covariates recommended by a group of estate agents, decides to use a multilayer perceptron as defined below in the `sklearn` package. The researcher has access to a total of 200 data points and chooses a random subset of 150 for training, reserving the remaining 50 for a final evaluation of the model.

```
MLPClassifier(hidden_layer_sizes=1000,
              activation='relu', *,
  solver='adam',
  alpha=0.0001,
  batch_size='auto',
  learning_rate='constant',
  learning_rate_init=0.001,
  power_t=0.5,
  max_iter=200000,
  shuffle=True,
  random_state=None,
  tol=0.0001,
  verbose=False,
  warm_start=False,
  momentum=0.9,
  nesterovs_momentum=True,
  early_stopping=False,
  validation_fraction=0.1,
  beta_1=0.9,
  beta_2=0.999,
  epsilon=1e-08,
  n_iter_no_change=10,
  max_fun=15000)
```

Which of the following statements is true?

1. ✓ It is possible that the error on the training set reduces to a negligibly small value.
2. ✓ The average error on the 50 test samples is likely to be disappointingly high.

3. If the model does not generalize, additional hidden layers should be used.

4. I would advice against setting `early_stopping=False`

5. Your supervisor has asked you to apply Expectation-Maximisation for clustering. Which of the following statements is true?

1. The objective is guaranteed to converge to a global optimum.
2. Convergence rate depends on the properties of the likelihood function.
3. The log-likelihood, $\ln p_{\boldsymbol{\theta}}(\mathbf{X})$ decreases monotonically.

6. Consider Expectation-Maximisation (EM) for Gaussian Mixture Models (GMMs). Assume the GMM has K components and that all GMM components share the same covariance $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_K = \epsilon\mathbf{I}$, where $\mathbf{I}$ denotes the identity matrix and ϵ is a constant. Which of the following statements is true?

1. EM approaches the k-means solution if $\epsilon = 1$.
2. EM approaches the k-means solution if $\epsilon \to \infty$.
3. EM approaches the k-means solution if $\epsilon \to 0$.