02471 Machine Learning for Signal Processing

Problem set 1

This problem set is based on the teaching material from week 1 and 2 as well as central knowledge assumed to be known beforehand.

This problem set will be a passed/failed. All problems are weighted equally (regardless of number of sub-questions) and you must obtain a score of 50% to pass.

1.1 Cross validation

Which one of the following statements pertaining to cross-validation is correct?

(a) Leave-one-out cross-validation is computationally expensive since as many models as observations needs to be trained.

(b) For datasets with very few observations it is in general worse to use leave-one-out cross-validation rather than 5-fold cross-validation.

(c) Two levels of cross-validation is necessary in order to determine the optimal set of parameters for a model.

(d) 4-fold cross-validation is the same as the hold-out method when 25% is held out.

Please justify your answer in 3–5 lines.

1.2 Double-folded cross validation

Alice is considering a linear regression model for a dataset comprised of $N = 5000$ observations. She wishes to both select the optimal regularization strength as well as estimate the generalization error of the model at the optimal regularization strength. To simplify the problem, she only considers the following 4 possible values of the regularization strength λ:

$$\lambda \in \{10^{-2}, 10^{-1}, 10^{0}, 10^{1}\}$$

Alice opts for a two-level strategy in which she uses the hold-out method to estimate the generalization error and cross-validation is used to select the optimal regularization strength, i.e. the dataset is first divided into a validation set $D_{validation}$, comprised of 10% of the full dataset, and the remainder D_{CV} is used for cross-validation. Alice uses standard K = 5 fold cross-validation to select the optimal regularization strength on D_{CV} and, having estimated the optimal regularization strength, uses the hold-out method on D_{CV} and $D_{validation}$ to estimate the generalization error.

Suppose for any fixed value of the regularization strength, the time taken to train the weights of the linear regression model on a dataset of size N_{train} is N_{train}^2 units of time and the time taken to test a trained model on a dataset of size N_{test} is $1/2N_{test}^2$ units of time. Suppose the duration of all other tasks is negligible, what is the total time taken for the entire procedure?

Provide the result in the units of time. Include the calculations.

1.3 k-nearest neighbor classification

In this question we consider a dataset that is created using queries where a query will return a set of numbers that is organised as a vector. To classify the result returned by the query, we will use a k-nearest neighbor classifier based on the Euclidean distance between the results given in the following table:

	R1	R2	R3	B1	B2	B3
R1	0.00	3.80	2.84	2.91	1.59	2.61
R2	3.80	0.00	1.05	6.01	4.42	3.82
R3	2.84	1.05	0.00	4.97	3.37	2.86
B1	2.91	6.01	4.97	0.00	1.61	2.62
B2	1.59	4.42	3.37	1.61	0.00	1.51
B3	2.61	3.82	2.86	2.62	1.51	0.00

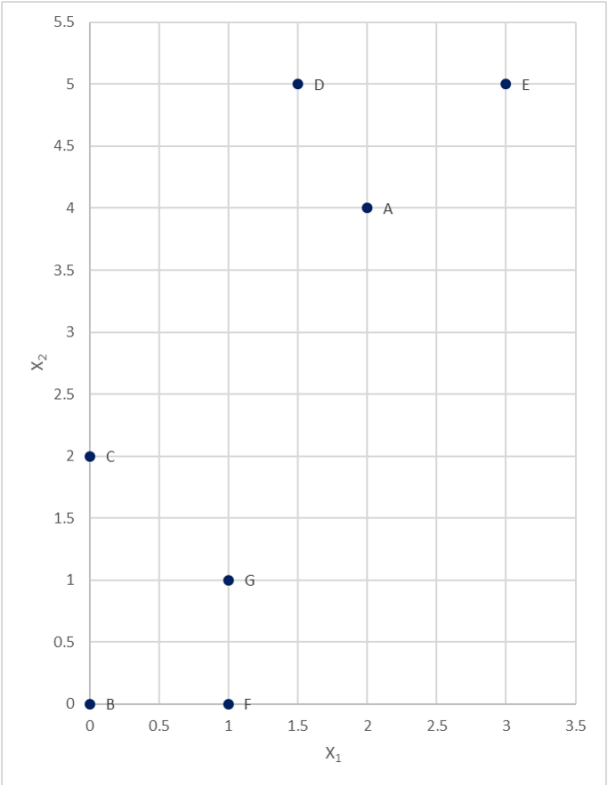
We will classify each of results from the queries in succession and determine whether the right queries returned a class one object (given in red i.e. R1, R2, R3) or class two object (given in blue, i.e. B1, B2, B3). When classifying we will use k = 3. The analysis is based only on the data given in the table.

Specify which points are classified correctly.

1.4 k-means clustering

Consider simple 2-dimensional data set comprised of N = 7 observations as shown in the table below.

i	X_1	X_2
A	2	4
B	0	0
C	0	2
D	1.5	5
E	3	5
F	1	0
G	1	1



Suppose we wish to apply K-means clustering to the data set and the K = 2 two-dimensional cluster centers are initialized in $\mu_1 = (2, 0.5)$ and $\mu_2 = (1.5, 3.5)$.

Question 1.4.1

How many points belong to cluster one and two respectively at initialization?

Question 1.4.2

After one iteration of the K-means clustering algorithm, what is the cluster center of μ_1?

1.5 Expectation is a linear operator

Show that the expectation operator is a linear operator by using the definition of the expectation for continous random variables. A linear operator fullfills

$$\mathbb{E}[a \cdot \mathrm{x} + b \cdot \mathrm{y}] = a\mathbb{E}[\mathrm{x}] + b\mathbb{E}[\mathrm{y}]$$

where a and b er deterministic scalars, x and y are continous random variables.

1.6 Discrete expectation

In the casino game of roulette, a wheel is spun, and a little ball drops into one of many numbered spots. In this question we consider an American roulette wheel as shown below:

The simplest way to bet is either on "red" or "black". Let's say you bet ("ante up") a dollar on red. If the wheel is spun, and it comes up red, you win \$2 (which means you profited \$1). If not, you lose the \$1.

What can you expect to earn in the long run if you bet this way on an American roulette wheel? Give your answer as earnings per bet with 2 significant digits. Please include the calculations.

1.7 Probabilities

We, the engineers, are much smarter than doctors. What better way to demonstrate your intellectual superiority than by answering something that stumps doctors. A few years back there was an enlightening study on how doctors miscalculate. Doctors were asked to answer a question that they would have to answer in their everyday routine: "If a woman has a positive mammogram, what is the probability that she has breast cancer?" The relevant statistics you need to know are as follows:

- The probability that a woman has breast cancer is 0.7%
- If a woman has breast cancer, there is a probability of 90% that the mammogram will be positive.
- If she has no breast cancer the probability of a positive mammogram is 8%

Find the probability that a patient has breast cancer given a positive mammogram.

Give your answer as a probability with two significant digits.

1.8 Convolutions

Consider the following two digital signals.

$$x(n) = \begin{cases} \frac{1}{3}n & \text{for } 0 \leq n \leq 6 \\ 0 & \text{elsewhere} \end{cases}$$

$$h(n) = \begin{cases} 1 & \text{for } -2 \leq n \leq 2 \\ 0 & \text{elsewhere} \end{cases}$$

Determine the convolution $y(n) = x(n) * h(n)$ of the two signals, that is, calculate $y(n)$ for each value of n. Include the calculation.

1.9 Spectrum of a sinusoid

In this question we will calculate the spectrum of a continuous-time sinusoid. Solve the two questions below:

Question 1.9.1

Write out the analog sinusoid $x(t) = A\cos(2\pi F_0 t + \theta)$, $-\infty < t < \infty$ as a complex exponential signal.

Question 1.9.2

What are the Fourier coefficients for the analog sinusoid? (hint: sketch the magnitude and phase spectrum)

1.10 Principal component analysis

A principal component analysis is carried out on a data based on $x_1, \cdots, x_4$. The mean is subtracted from each attribute and the singular value decomposition (SVD) is applied to the data matrix of size 150x4. From the SVD we obtain for the matrices S and V:

$$S = \begin{pmatrix} 95.95 & 0.00 & 0.00 & 0.00 \\ 0.00 & 17.76 & 0.00 & 0.00 \\ 0.00 & 0.00 & 3.46 & 0.00 \\ 0.00 & 0.00 & 0.00 & 1.88 \end{pmatrix}$$

$$V = \begin{pmatrix} -0.75 & -0.38 & -0.51 & -0.17 \\ 0.28 & 0.55 & -0.71 & -0.34 \\ 0.50 & -0.68 & -0.06 & -0.54 \\ 0.32 & -0.32 & -0.48 & 0.75 \end{pmatrix}$$

We note that both S and V above have been rounded to the first couple of significant digits.

Compute the explained variance of the first two components. Give your answer with 2 significant digits.

1.11 Lagrange multipliers

This exercise will formulate PCA as an optimization problem, and show how to use Lagrangian multipliers to arrive at the PCA solution. If you are not familiar with PCA, be sure to skim the first part of section 19.3 before completing this question.

Let $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, ..., $\boldsymbol{x}_n$ be a collection of vectors. For all n, consider:

$$z_{1,n} = \boldsymbol{u}_1^T \boldsymbol{x}_n$$

that is, $z_{1,n}$ is the scalar projection of $\boldsymbol{x}_n$ onto the vector $\boldsymbol{u}_1$, where we constrained the vector $\boldsymbol{u}_1$ to be a unit vector, i.e $\boldsymbol{u}_1^T \boldsymbol{u}_1 = 1$. The vector $\boldsymbol{z}_1$ containing all the $z_{1,n}$ can be written as:

$$\boldsymbol{z_1} = \boldsymbol{u}_1^T \mathbf{x}$$

where $\mathbf{x}$ is the matrix whose columns are the $\boldsymbol{x}_n$. PCA identifies the direction of $\boldsymbol{u}_1$ such that $\boldsymbol{z}_1$ has maximum variance.

Question 1.11.1

Using the expression for the sample variance and assuming the vectors have zero mean, show that $\text{var}[\boldsymbol{z}_1] = \boldsymbol{u}_1^T \hat{\Sigma}_x \boldsymbol{u}_1$ where $\hat{\Sigma}_x$ is the sample covariance matrix of $\mathbf{x}$.

Question 1.11.2

We can now construct our optimization problem

$$\begin{aligned} \boldsymbol{u}_1 &= \arg\max_{\boldsymbol{u}} \ \boldsymbol{u}^T \hat{\Sigma}_x \boldsymbol{u} \\ \text{s.t.} \quad & \boldsymbol{u}^T \boldsymbol{u} = 1 \end{aligned}$$

Now use Lagrangian multipliers to rewrite the optimization problem to a function, $L(\boldsymbol{u}, \lambda)$, and show that we arrive at an eigenvalue problem. Argue that our PCA optimization problem is now solved by selecting the $\boldsymbol{u}$ with the highest eigenvalue, λ, as the solution, thus having that $\boldsymbol{u}_{\lambda_{\max}}$ is the first principal component.