02471 Machine Learning for Signal Processing - Fall 24

Problem set 1

Name	Student ID
Jakob Ketmig	s194264

December 8, 2024

Contents

1 Cross validation (1.1)

The answer option (a) is correct. Leave-out-one cross-validation (LOO-CV) uses each data point once as a validation set, while the remaining $n-1$ observations are used for training. Hence, you train n models and the required computational budget is proportional to the size of your dataset N, making it infeasible on larger ones.

2 Double-folded cross validation (1.2)

From the text of the problem Alice has supplied us with the following known quantities:

$$D = 5000\ ,\ D_{val} = 500\ ,\ D_{CV} = 4500\ ,\ \lambda \in \{10^{-2}, 10^{-1}, 10^{0}, 10^{1}\}\ ,\ K = 5,$$

where λ represents the regularization strength of the linear regression model on a dataset, which has been split into validation D_{val} and cross-validation D_{CV} sets, respectively, using $K = 5$ folds. Given the assumptions stated in the problem of the model's training and testing time, we have the following split of data for each of the K folds:

$$D_{train} = \frac{K-1}{K} \times D_{CV} = \frac{4}{5} \times 4500 = 3600 \tag{1}$$

$$D_{test} = D_{CV} - D_{train} = 4500 - 3600 = 900 \tag{2}$$

Calculating the units of time spent training and testing, respectively, for a singular fold in the cross-validation yields the following units of time t spent:

$$T_{train}^{k=1} = D_{train}^2 = 3600^2 = 12.960.000t \tag{3}$$

$$T_{test}^{k=1} = \frac{1}{2} D_{test}^2 = \frac{1}{2} \times 900^2 = 405.000t \tag{4}$$

This in turn means a total of $K = 5$ folds with 4 different regularization strengths $|\lambda|$ would require:

$$T_K = |\lambda| \times K \times (T_{train}^{k=1} + T_{test}^{k=1}) = 4 \times 5 \times (12.960.000t + 405.000t) = 267.300.000t \tag{5}$$

After training Alice finds the optimal λ^*, which she trains her final model on using the **full** D_{CV}^{full} and tests on the D_{test}^{full}. Again, this would in turn yield a training and test time of:

$$N_{CV}^{full} = (D_{CV}^{full})^2 = 4500^2 = 20.250.000t \tag{6}$$

$$N_{test}^{full} = \frac{1}{2}(D_{test}^{full})^2 = \frac{1}{2} \times 500^2 = 125.000t \tag{7}$$

Putting it all together for the CV and the final model:

$$T_{total} = T_K + N_{CV}^{full} + N_{test}^{full} = (267.300.000 + 20.250.000 + 125.000)t = 287.675.000t \tag{8}$$

3 KNN Classification (1.3)

Given the dataset considered which used a Euclidean distance, we will run the KNN algorithm using $k = 3$ and determine which points are classified correctly into the blue and red classes. Reading off the table at **R1**, it is evident that its closest neighbours are **B2**$= 1.59$, **B3**$= 2.61$, and **R3**$= 2.84$. Now, using majority voting, it is equally evident that this point is classified as blue using the algorithm.

An overview is provided in table 1 where the same procedure is executed and the results are listed, however the conclusion is that all but **R1** are classified correctly.

Node	Nearest Neighbours ($k = 3$)	Classified as
R1	B2, R3, B3	blue
R2	R3, B3, R1	red
R3	R2, R1, B3	red
B1	B2, B3, R1	blue
B2	B3, R1, B1	blue
B3	B2, R1, B1	blue

Table 1: Classification Results for KNN ($k = 3$)

4 K-means clustering (1.4)

Considering the data supplied we will use $K = 2$ for our clustering technique initialized at $\mu_1 = (2, 0.5)$ and $\mu_2 = (1.5, 3.5)$.

4.1 Part 1

At initialization we have the following distribution of points between the two points using an Euclidean distance:

$$d_{i,j} = \sqrt{(x_i - \mu_{j,x})^2 + (y_i - \mu_{j,y})^2}, \tag{9}$$

where $\mu_{j,x}$ and $\mu_{j,y}$ are the coordinates at the j-th center, respectively, while (x_i, y_i) are the coordinates of the i-th observation in the dataset. This yields the following for $\mu_1 = (2, 0.5)$ using two significant digits:

$$\begin{aligned}
d_{A,\mu_1} &= \sqrt{(2-2)^2 + (4-0.5)^2} = \sqrt{0 - (3.5)^2} = 3.50 \\
d_{B,\mu_1} &= \sqrt{(0-2)^2 + (0-0.5)^2} = \sqrt{4.25} \approx 2.06 \\
d_{C,\mu_1} &= \sqrt{(0-2)^2 + (2-0.5)^2} = \sqrt{6.25} \approx 2.69 \\
d_{D,\mu_1} &= \sqrt{(1.5-2)^2 + (5-0.5)^2} = \sqrt{20.5} \approx 4.51 \\
d_{E,\mu_1} &= \sqrt{(3-2)^2 + (5-0.5)^2} = \sqrt{21.25} \approx 4.58 \\
d_{F,\mu_1} &= \sqrt{(1-2)^2 + (0-0.5)^2} = \sqrt{1.25} \approx 1.12 \\
d_{G,\mu_1} &= \sqrt{(1-2)^2 + (1-0.5)^2} = \sqrt{1.25} \approx 1.12
\end{aligned}$$

And correspondingly for $\mu_2 = (1.5, 3.5)$:

$$\begin{aligned}
d_{A,\mu_2} &= \sqrt{(2-1.5)^2 + (4-3.5)^2} = \sqrt{0.50} \approx 0.71 \\
d_{B,\mu_2} &= \sqrt{(0-1.5)^2 + (0-3.5)^2} = \sqrt{14.50} \approx 3.87 \\
d_{C,\mu_2} &= \sqrt{(0-1.5)^2 + (2-3.5)^2} = \sqrt{4.50} \approx 2.12 \\
d_{D,\mu_2} &= \sqrt{(1.5-1.5)^2 + (5-3.5)^2} = \sqrt{2.25} \approx 1.50 \\
d_{E,\mu_2} &= \sqrt{(3-1.5)^2 + (5-3.5)^2} = \sqrt{4.50} \approx 2.12 \\
d_{F,\mu_2} &= \sqrt{(1-1.5)^2 + (0-3.5)^2} = \sqrt{12.50} \approx 3.50 \\
d_{G,\mu_2} &= \sqrt{(1-1.5)^2 + (1-3.5)^2} = \sqrt{6.50} \approx 2.58
\end{aligned}$$

Hence, it is evident that the points $\{B, F, G\}$ are assigned to μ_1 while $\{A, C, D, E\}$ are assigned to μ_2. The answer is therefore 3 and 4, respectively.

4.2 Part 2

After one iteration, the cluster center of μ_1 can be found as the mean of the x and y coordinates of the points assigned to the cluster, ie. $\{B, F, G\}$:

$$\mu_1^{'} = (x_{avg}^{\mu_1}, y_{avg}^{\mu_1}) = \left(\frac{0+1+1}{3}, \frac{0+0+1}{3}\right) = \left(\frac{2}{3}, \frac{1}{3}\right) \tag{10}$$

5 $\mathbb{E}$ is a linear operator (1.5)

The definition of an expectation for a continuous random variable z is given by integrating over the entirety of the probability density function $f(z)$:

$$\mathbb{E}[z] = \int_{-\infty}^{\infty} a \cdot f(z) \, \mathrm{d}x \tag{11}$$

Considering the expression $z = ax + by$, we have:

$$\begin{aligned}
\mathbb{E}[ax + by] &= \int_{-\infty}^{\infty} (ax + by) \cdot f(x, y) \, \mathrm{d}x \, \mathrm{d}y && (12)\\
&= a \int_{-\infty}^{\infty} x \cdot f(x, y) \, \mathrm{d}x \, \mathrm{d}y + b \int_{-\infty}^{\infty} y \cdot f(x, y) \, \mathrm{d}x \, \mathrm{d}y && (13)\\
&= a \int_{-\infty}^{\infty} x \cdot f(x) \, \mathrm{d}x \, \mathrm{d}y + b \int_{-\infty}^{\infty} y \cdot f(y) \, \mathrm{d}x \, \mathrm{d}y && (14)\\
&= a\mathbb{E}[x] + b\mathbb{E}[y] && (15)
\end{aligned}$$

From (13) to (14) we use that integration is a linear operator to break up the integral. Furthermore, at line (14) note that $f(x, y)$ is a joint distribution, however we may marginalize for each variable, respectively, as it only depends on x and y respectively, yielding the result. Hence, the expectation operator is indeed linear.

6 Discrete Expectation (1.6)

There are a total of 38 tiles in the wheel distributed as $n_{red} = n_{black} = 18$ and $n_{green} = 2$. The expected value per bet on red, or earnings, can be evaluated using two significant digits as the following:

$$\begin{aligned}
EVPB &= (P_{win} \times \$_{win}) + (P_{loss} \times \$_{loss}) = \left(\frac{n_{red}}{N} \times \$_{win}\right) + \left(\frac{\neg n_{red}}{N} \times \$_{loss}\right) && (16)\\
&= \left(\frac{18}{38} \times 1\right) + \left(\frac{18+2}{38} \times (-1)\right) = \frac{18-20}{38} = \frac{-2}{38} \approx -0.053 && (17)
\end{aligned}$$

7 Probabilities (1.7)

Given the information in the problem formulation, it seems evident we have to use Bayes' theorem. Writing up the information, where BC means has breast cancer, M_p means positive mammogram and negations $\neg$ correspond to the opposite, of course. The known facts are:

- $P(BC) = 7\%$: Probability of having breast cancer. Hence, $P(\neg BC) = 100\% - 7\% = 99.3\%$.
- $P(BC|M_p) = 90\%$: Breast cancer if positive mammogram.
- $P(M_p|\neg BC) = 8\%$: No breast cancer, but positive mammogram.

Hence, using Bayes' theorem:

$$P(BC|M_p) = \frac{P(BC|M_p)P(BC)}{P(M_p)} \tag{18}$$

First, we must find $P(M_p)$ in order to calculate the above. This is done by using the law of total probability, ie. we can express it as:

$$\begin{aligned} P(M_p) &= P(M_p|BC)P(BC) + P(M_p|\neg BC)P(\neg BC) \\ &= (0.9 \times 0.07) + (0.993 \times 0.08) = 0.063 + 0.07944 = 0.14244 \end{aligned}$$

Inserting into (18) we have:

$$P(BC|M_p) = \frac{0.9 \times 0.07}{0.14244} = \frac{0.063}{0.14244} \approx 0.4423 = 44.23\% \tag{19}$$

8 Convolutions (1.8)

Considering the signals supplied in the problem, we have to determine $y(n) = x(n) * h(n)$. Here we will use the discrete-time convolution formula $y(n) = \sum_{k=-\infty}^{\infty} x(k)h(n-k)$. Explicitly, for the signals we have:

$$\begin{aligned} x(n) &= \{0, \frac{1}{3}, \frac{2}{3}, 1, \frac{4}{3}, \frac{5}{3}, 2\} \text{ for } 0 \leq n \leq 6 \\ h(n) &= 1 \text{ for } -2 \leq n \leq 2 \end{aligned}$$

Recall, the signal h is slid over the range of x at each position n. The length of y is influenced by the total span of both x and h, ie. from $y_{lower} = x_{lower} + h_{lower} = 0 + (-2) = -2$ to $y_{upper} = x_{upper} + h_{upper} = 6 + 2 = 8$. Inserting the values for $y(n)$ yields a convolved signal with values:

$$\begin{aligned} y(-2) &= \sum_{k=-\infty}^{\infty} x(k)h(-2-k) = 0 \;\; (x(k) = 0 \,\forall\, k \leq 0) \\ y(-1) &= 0 \\ y(0) &= x(0)h(0) = 0 \times 1 = 0 \\ y(1) &= x(0)h(1-0) + x(1)h(1-1) = 0 \times 1 + \frac{1}{3} \times 1 = \frac{1}{3} \\ y(2) &= x(0)h(2) + x(1)h(1) + x(2)h(0) = 0 + \frac{1}{3} + \frac{2}{3} = 1 \\ y(3) &= x(1)h(2) + x(2)h(1) + x(3)h(0) = \frac{1}{3} + \frac{2}{3} + 1 = 2 \\ y(4) &= x(2)h(2) + x(3)h(1) + x(4)h(0) = \frac{2}{3} + 1 + \frac{4}{3} = 3 \\ y(5) &= x(3)h(2) + x(4)h(1) + x(5)h(0) = 1 + \frac{4}{3} + \frac{5}{3} = 4 \\ y(6) &= x(4)h(2) + x(5)h(1) + x(6)h(0) = \frac{4}{3} + \frac{5}{3} + 2 = 5 \\ y(7) &= x(5)h(2) + x(6)h(1) = \frac{5}{3} + 2 = \frac{11}{3} \\ y(8) &= x(6)h(2) = 2 \end{aligned}$$

$y(n)$ is 0 elsewhere besides the interval $n \in [0, 8]$ as shown above. Some, coincidentally, are also zero.

9 Spectrum of a sinusoid (1.9)

9.1 Analog to complex exponential

Using Euler's formula we have:

$$cos(\theta) = \frac{e^{j\theta} + e^{-j\theta}}{2} \tag{20}$$

Hence, we can re-write the analog sinusoid as:

$$x(t) = A\cos(2\pi F_0 t + \theta) = A\left(\frac{e^{j(2\pi F_0 t+\theta)} + e^{-j(2\pi F_0 t+\theta)}}{2}\right) \tag{21}$$

$$= \frac{A}{2}\left(e^{j(2\pi F_0 t+\theta)} + e^{-j(2\pi F_0 t+\theta)}\right) = \frac{A}{2}e^{j\theta}e^{j2\pi F_0 t} + \frac{A}{2}e^{-j\theta}e^{-j2\pi F_0 t} \tag{22}$$

So, we have two complex exponential components at F_0 and $-F_0$, which aligns with the information presented in the lecture. This means the Fourier spectrum has impulses at these two frequencies, which correspond to non-zero Fourier coefficients.

9.2 Fourier coefficients

In Fourier series the periodic signal is represented as:

$$x(t) = \sum_{n=-\infty}^{\infty} c_n e^{j2\pi n F_0 t} \tag{23}$$

The coefficients are directly visible from (22), meaning:

$$c_1 = \frac{A}{2}e^{j\theta} \tag{24}$$

$$c_{-1} = \frac{A}{2}e^{-j\theta} \tag{25}$$

$$c_n = 0 \ \textbf{elsewhere} \tag{26}$$

10 PCA (1.10)

In order to find the exlained variance of the first two principal components, we use the matrix $\mathbf{S}$ with singular values σ_i. Recall, each entry corresponds to the square root of the eigenvalues of the covariance matrix of the data $\mathbf{X}$, hence relating each principal component's captured variance. The total variance can be found to be $\text{Tr}(\mathbf{S}) = \sigma_i^2$. Hence, the explained variance of the first two principal components can be calculated as:

$$\text{Variance Explained}(PC1, PC2) = \frac{\sigma_1^2 + \sigma_2^2}{\sum_{i=1}^{4}\sigma_i^2} = \frac{95.95^2 + 17.76^2}{95.95^2 + 17.76^2 + 3.46^2 + 1.88^2} = \frac{9523.02}{9538.53} \approx 0.998 \tag{27}$$

So the first two principal components explain 99.8% of the variance in the dataset.

11 Lagrange Multipliers (1.11)

11.1 Variance expression proof

We are to show that $\text{var}[\mathbf{z}_1] = \mathbf{u}_1^T\hat{\Sigma}_x\mathbf{u}_1$, where $\hat{\Sigma}_x$ is the sample covariance. The variance of $\mathbf{z}_1$ is essentially the variance of the projected data onto $\mathbf{u}_1$. By the definition of variance, we have:

$$\text{var}[\mathbf{z}_1] = \mathbb{E}[(\mathbf{z}_1 - \mathbb{E}[\mathbf{z}_1])^2] = \mathbb{E}[\mathbf{z}_1^2] \text{ (centered data, ie. } \mathbb{E}[\mathbf{z}_1] = 0) \tag{28}$$

$$= \mathbb{E}[(\mathbf{u}_1^T x)^2] = \mathbf{u}_1^T \mathbb{E}[xx^T]\mathbf{u}_1 = \mathbf{u}_1^T \hat{\Sigma}_x \mathbf{u}_1 \tag{29}$$

11.2 Proof and relate to PCA

Under the constraint that $\mathbf{u}_1$ is a unit vector, ie. $\mathbf{u}_1^T\mathbf{u}_1 = 1$, we try to optimize w.r.t. $\mathbf{u}$ for $\mathbf{u}_1 = \arg\max_{\mathbf{u}} \text{var}[\mathbf{z}_1]$ by reformulating it using the general form of a Lagrange multiplier[1], which will be dependent on the eigenvalue λ and $\mathbf{u}$. This is given by $L(u, \lambda) = f(u) + \lambda \cdot g(u)$, where f is the objective function and g is the constraint. Rearranging the constraint and inserting we then have:

$$L(u, \lambda) = \mathbf{u}^T \hat{\Sigma}_x \mathbf{u} + \lambda \cdot (\mathbf{u}^T\mathbf{u} - 1) \tag{30}$$

Finding the stationary points of L on f subject to g, ie. taking the partial derivative w.r.t. u and setting it equal to zero:

$$\frac{\partial L}{\partial u} = 2\mathbf{u}\hat{\Sigma}_x - 2\lambda\mathbf{u} = 0 \Leftrightarrow \hat{\Sigma}_x\mathbf{u} = \lambda\mathbf{u} \tag{31}$$

This is the eigenvalue problem $Ax = \lambda x$, ie. the vector $\mathbf{u}$ that maximizes the objective is the eigenvector corresponding to the largest eigenvalue of the sampled variance of the data, which one may recall is indeed ordered in descending fashion w.r.t. λ. Thus the first principal component is given by $\mathbf{u}_{\lambda_{max}}$.

[1] The basic idea is to convert a constrained problem into a form such that the derivative test of an unconstrained problem can still be applied.

References