02471 Machine Learning for Signal Processing

Written examination: December 6, 2023.

Course name: Machine Learning for Signal Processing.

Course number: 02471.

Aids allowed: All DTU allowed aids are permitted.

Exam duration: 4 hours.

Weighting: The weighting is indicated in parentheses for each sub-problem.

This exam has 7 problems with a total of 18 questions, for a total of 100% weighting.

Hand-in: Hand-in on paper and/or upload a PDF file. Do not hand in duplicate information.

All answers must include relevant considerations and/or calculations/derivations.

It should be clear what theories and formulas were used from the curriculum.

Multiple choice: The following problems are multiple choice:

Problem 6.1.

If a multiple choice problem is answered correctly, you are given full points. For a wrong answer, you are subtracted 1/4 of the full point value. E.g, if a problem gives 5 points for a correct answer, a wrong answer will result in subtracting 1.25 points.

Problem 1 Parameter Estimation (15% total weighting)

In this problem we will consider the following data

n	0	1	2	3
y_n	0.5	-1.2	2.0	4.1

Problem 1.1 (5% weighting)

Consider polynomial regression using the squared error as the loss function and assume i.i.d noise. For modeling the response y_n, we will use a polynomial model of the form $f_n = \theta_0 + \theta_1 \cdot n + \theta_2 \cdot n^2$.

Determine the parameters $\hat{\boldsymbol{\theta}} = [\theta_0\ \theta_1\ \theta_2]^T$ of the model that minimizes the error. Describe the process for calculating this estimate, including the relevant matrices and vectors.

Problem 1.2 (5% weighting)

Using the same data, we consider a linear model on the form $f_n = \theta_0 + \theta_1 \cdot n$, and we assume non-white Gaussian noise, where the variance of the noise is 1, and the covariance of the noise between two successive samples is 0.2.

Determine the parameters $\hat{\boldsymbol{\theta}} = [\theta_0\ \theta_1]^T$ of the model that minimizes the squared error using this noise assumption. Write the relevant matrices and vectors used to estimate $\hat{\boldsymbol{\theta}}$.

Problem 1.3 (5% weighting)

We now consider the case of general parameter estimation. Assume that your parameter is estimated using 10 different datasets (realizations). For each dataset, we have an unbiased estimator, denoted $\hat{\theta}_i$, where the variance of these individual estimators are $\sigma^2 = 1$. We aggregate the estimates using:

$$\hat{\theta}_{\text{agg}} = \frac{1}{10} \sum_{i=1}^{10} \hat{\theta}_i$$

Compute the variance of the estimator $\hat{\theta}_{\text{agg}}$.

Specify formulas, the numerical value, and assumptions made.

Problem 2 Sparse learning (10% total weighting)

In this problem, we consider sparse learning and apply ℓ_1 regularization and linear regression:

$$\arg\min_{\boldsymbol{\theta}} \left\{ \|\boldsymbol{y} - X^T\boldsymbol{\theta}\|^2 + \lambda \|\boldsymbol{\theta}\|_1 \right\}$$

Problem 2.1 (5% weighting)

For training, we will use the "iterative shrinkage/thresholding" scheme:

$$\boldsymbol{\theta}^{(i)} = S_{\lambda\mu}\left(\boldsymbol{\theta}^{(i-1)} + \mu X^T \boldsymbol{e}^{(i-1)}\right)$$

where $S_{\lambda\mu}(\cdot)$ denotes the shrinkage/thresholding function.

We run one iteration of the algorithm. Assume the following:

$$\boldsymbol{\theta}^{(i-1)} = \begin{bmatrix}1 & 0.9\end{bmatrix}^T$$
$$X^T \boldsymbol{e}^{(i-1)} = \begin{bmatrix}-0.5 & -0.3\end{bmatrix}^T$$

Determine the smallest value of λ that results in both components in $\boldsymbol{\theta}^{(i)}$ being zero with $\mu = 0.1$.

Problem 2.2 (5% weighting)

Assume we have chosen $\lambda = 1$, and we obtain an reliable estimate of the noise of the data, denoted $\sigma^2 = 1$. How does the original regression problem relate to the prior distribution of $\boldsymbol{\theta}$? Determine the parameters of the prior distribution of $\boldsymbol{\theta}$.

Problem 3 Linear filtering (30% total weighting)

In this problem we are considering the echo canceling setup, where we assume the echo canceller is a linear FIR filter, with l coefficients. The setup is depicted in Figure 1, page 3.

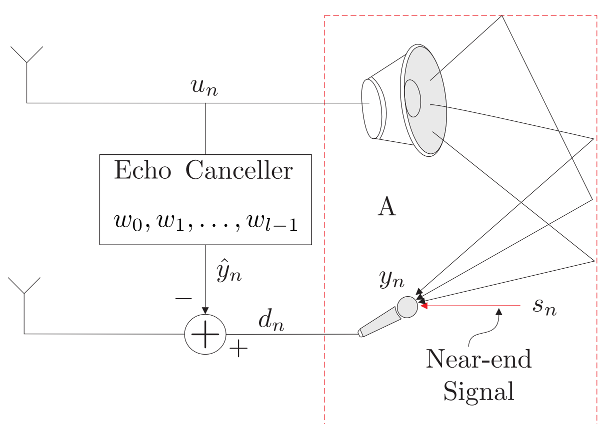

Figure 1: Problem 3 setup.

Assume that we have a u_n sequence as follows:

n	0	1	2	3	4	5	6
u_n	1	6	5	2	3	1	6

Problem 3.1 (5% weighting)

The filter has the coefficients $\boldsymbol{w} = \begin{bmatrix} 0.50 & 0.30 & w_2 \end{bmatrix}^T$.

The output of the filter at $n = 4$ is $\hat{y}_4 = 3.1$. Determine the numerical value of w_2 that was used to compute $\hat{y}_4$.

Problem 3.2 (7% weighting)

We will now consider all signals as wide-sense stationary stochastic processes.

We are informed that room A has a room impulse response of $H_A = [0, h_1, h_2]$, such that the proportion of u_n that is received by the microphone y_n is H_A convolved with $\mathbf{u}_n$ (where $\mathbf{u}_n = \begin{bmatrix} \mathrm{u}_n & \mathrm{u}_{n-1} & \mathrm{u}_{n-2} \end{bmatrix}^T$).

Derive an analytical expression for the cross-correlation function $r_{yu}(k)$.

Problem 3.3 (5% weighting)

We now obtain precise measurements of the following correlation functions, where $r_u(k)$ denotes the correlation function for the signal u_n, and $r_{du}(k)$ denotes the cross-correlation function between d_n and u_n:

k	0	1	2	3
$r_u(k)$	1.0	0.9	0.7	0.5
$r_{du}(k)$	0.45	0.55	0.30	0.20

Determine the filter coefficient values $\boldsymbol{w}$ (still with a filter length $l = 3$) based on the measured correlation values.

Additionally, we need a filter which has, at most, a minimum mean squared error (MMSE) of 0.1. Determine the highest tolerable value for $r_d(0)$.

You should specify both the exact formulas and the numerical values.

Problem 3.4 (5% weighting)

We now have a change of the signal u_n with new correlation values, and decide to learn the filter weights using the LMS algorithm. We are given the following knowledge of the covariance matrix of u_n where u_n is assumed to be a wide-sense stationary stochastic process.

$$\Sigma_u = \begin{bmatrix} 2.00 & 1.40 & 0.70 \\ 1.40 & 2.00 & 1.40 \\ 0.70 & 1.40 & 2.00 \end{bmatrix}$$

We want to choose the step size such that we are sure that the filter converges and the weight estimator has bounded variance.

Determine the maximum value for the step-size μ.

You should specify both the exact formulas and the numerical values.

Problem 3.5 (8% weighting)

We now assume that the time-varying component of the entire system is adequately modeled using the following model:

$$\begin{aligned} \mathrm{d}_n &= \boldsymbol{w}_{o,n-1}^T \mathbf{u}_n + \eta_n \\ \boldsymbol{w}_{o,n} &= \boldsymbol{w}_{o,n-1} + \omega_n \\ \mathbb{E}\left[\omega_n \omega_n^T\right] &= \begin{bmatrix} 0.03 & 0.00 & 0.00 \\ 0.00 & 0.02 & 0.00 \\ 0.00 & 0.00 & 0.01 \end{bmatrix} \end{aligned}$$

Where η_n and ω_n are assumed to follow zero-mean normal distributions, and u_n has the same covariance matrix as the previous problem.

We will use NLMS instead of LMS, and want to determine the step-size μ that ensures that the time-varying component of the excess MSE reaches at maximum 0.1. Additionally, NLMS must be guaranteed to remain stable.

Determine the range of values for μ that fulfill these requirements.

You should specify both the formulas and the numerical values.

Problem 4 Dictionary learning (10% total weighting)

This problem concerns Independent Component Analysis (ICA) using Mutual information.

Problem 4.1 (5% weighting)

Let us assume that we have two sources, s_1 and s_2, which are statistically independent, and two observable variables x_1 and x_2, defined as

$$\begin{aligned} \mathrm{x}_1 &= \mathrm{s}_1 + \mathrm{s}_2 \\ \mathrm{x}_2 &= \mathrm{s}_2 \end{aligned}$$

We will now unmix the signals x_1 and x_2 using ICA uising an unmixing matrix W:

$$\begin{bmatrix} \hat{\mathrm{s}}_1 \\ \hat{\mathrm{s}}_2 \end{bmatrix} = W \begin{bmatrix} \mathrm{x}_1 \\ \mathrm{x}_2 \end{bmatrix}$$

Identify W such that $\hat{\mathrm{s}}_1$ and $\hat{\mathrm{s}}_2$ has an average mutual information $I(\hat{\mathrm{s}}_1, \hat{\mathrm{s}}_2) = 0$.

Show formally that $I(\hat{\mathrm{s}}_1, \hat{\mathrm{s}}_2) = 0$ for your solution.

Problem 4.2 (5% weighting)

We know that the ICA model cannot identify the scale and direction of the identified vectors of the unmixing matrix W, and consequently in the mixing matrix A.

Write a short proof that shows the ICA model has a scaling ambiguity, that is, for a specific identified W (or A), this W (or A) can be scaled, vector-wise, and still result in sources that are the same, up to scaling factor.

Problem 5 Hidden Markov Models (12% total weighting)

This problem concerns a Hidden Markov Model (HMM) where x_n denotes the state of the chain at time n, and y_n denotes the observation at time n.

Problem 5.1 (5% weighting)
Suppose a 3-state HMM have the initial state probability vector

$$P_k = \begin{bmatrix} 0.25 & 0.25 & 0.50 \end{bmatrix}^T$$

and the following state transition probabilities

$$P_{ij} = \begin{bmatrix} 0.5 & 0.5 & 0.0 \\ 0.0 & 0.2 & 0.8 \\ 0.0 & 0.9 & 0.1 \end{bmatrix}$$

We will construct a state sequence of length 8, by using each of the following state sub-sequences exactly once:

$$A = \begin{bmatrix} 1 & 2 \end{bmatrix}, \quad B = \begin{bmatrix} 2 & 3 \end{bmatrix}, \quad C = \begin{bmatrix} 3 & 2 \end{bmatrix}, \quad D = \begin{bmatrix} 3 & 3 \end{bmatrix}$$

E.g., the composition 'ABCD' will give the state sequence

$$x_n = \begin{bmatrix} 1 & 2 & 2 & 3 & 3 & 2 & 3 & 3 \end{bmatrix}$$

Specify the composition that creates the most likely state sequence given the defined HMM. Justify your choices.

Problem 5.2 (7% weighting)
You are now informed that the 3-state HMM allows three possible observable actions, a_1, a_2, and a_3 with the following emission probabilities

	a_1	a_2	a_3
$P(y_n = a_i \mid x_n = 1)$	0.7	0.1	0.2
$P(y_n = a_i \mid x_n = 2)$	0.5	0.4	0.1
$P(y_n = a_i \mid x_n = 3)$	0.9	0.0	0.1

We want to compute the initial state probability vector P_k, instead of using the P_k from the original problem.

Based on observations, you additionally know that $P(y_1 = a_1) = 0.68$ and $P(y_1 = a_2) = 0.15$.

Compute the initial state probability vector P_k using the given information.

You should specify both the formulas and the numerical values.

Problem 6 Kalman filtering (8% total weighting)

In this problem we consider Kalman filtering.

Problem 6.1 (8% weighting)

Consider a one-dimensional discrete-time system being tracked using a Kalman filter. The system is defined by the following state-space equations:

$$\begin{aligned} \mathrm{x}_n &= \mathrm{x}_{n-1} + \eta_n \\ \mathrm{y}_n &= \mathrm{x}_n + \mathrm{v}_n \end{aligned}$$

Assume that η_n follows a zero-mean normal distribution with $\sigma_\eta^2 = 0.01$, and v_n follows a zero-mean normal distribution with $\sigma_v^2 = 0.1$. We now observe one observation, $y_1 = 2$.

Calculate the updated state estimate $\hat{x}_1$ and error covariance P_1 after receiving the first measurement. Use as initial state $x = 0$ and $P = 1$.

Select the correct statement:

A : $\hat{x}_1 = 1.818$, $P_1 = 0.101$
B : $\hat{x}_1 = 1.818$, $P_1 = 0.091$
C : $\hat{x}_1 = 1.980$, $P_1 = 0.010$
D : $\hat{x}_1 = 2.165$, $P_1 = 0.135$
E : $\hat{x}_1 = 1.523$, $P_1 = 0.241$
F : Don't know.

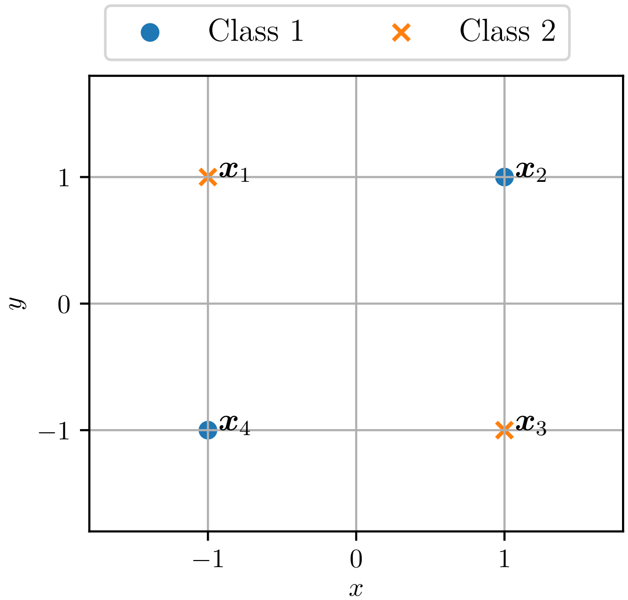

Figure 2: Problem 7.1.

Problem 7 Kernel methods (15% total weighting)

In this problem we will consider kernel methods.

Problem 7.1 (5% weighting)
In this problem we will use the Gaussian kernel

$$\kappa(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{y}\|^2}{2\sigma^2}\right)$$

We have observed the data points as depicted on Figure 2, page 8 as training data.

Write up the complete kernel matrix $\mathcal{K}$ for the data, where you have used the kernel function to evaluate the points. Use $\sigma = 2$.

You should specify both the formulas and the numerical values.

Problem 7.2 (5% weighting)
Now assume that two test points are given, $\boldsymbol{x}_1^{test}$ which belongs to class 1, and $\boldsymbol{x}_2^{test}$ which belongs to class 2. Their kernel values w.r.t. the training data (from Figure 2, page 8) are given as

	$\kappa(\cdot, \boldsymbol{x}_1)$	$\kappa(\cdot, \boldsymbol{x}_2)$	$\kappa(\cdot, \boldsymbol{x}_3)$	$\kappa(\cdot, \boldsymbol{x}_4)$
$\boldsymbol{x}_1^{test}$	0.97	0.59	0.46	0.75
$\boldsymbol{x}_2^{test}$	0.59	0.97	0.75	0.46

Use the representer theorem to determine how these two points can be correctly classified by finding suitable numerical values for θ_n, such that both are correctly classified using the same values for θ_n. Have a decision threshold as:

$$f(\boldsymbol{x}_1^{test}) > 0$$
$$f(\boldsymbol{x}_2^{test}) < 0$$

and have as many $\theta_n = 0$ as possible.

Problem 7.3 (5% weighting)
We consider a regression problem using support vector regression, fitted on five data points. We know that the data is fitted using $C = 1$ and $\epsilon = 0.1$, and the bias is estimated to $\hat{\theta}_0 = 0$. We have the following information:

n	1	2	3	4	5
$y_n - \hat{y}_n$	-0.14	0.04	0.11	-0.01	0.10
$\kappa(\boldsymbol{x}^{test}, \boldsymbol{x}_n)$	0.38	0.92	0.84	0.28	0.03

Computer either an exact value for $\hat{y}(\boldsymbol{x}^{test})$, or a tight bound for $\hat{y}(\boldsymbol{x}^{test})$, whichever is possible. You should specify both the formulas and the numerical values.