Deadline: Friday November 27th , 10:00 Please upload a `.pdf` file via LearnWeb. (e.g. exported Jupyter notebook)

1 Model Selection (8 points)

A [2p] Explain how one can detect whether a model is over- or underfitting.

B [2p] Explain how one can deal with a model that's over- or underfitting.

C [2p] Consider a binary classification problem where each class is generated by a Normal distribution.

- 50% of the datapoints belong to class A and are distributed as $p(x \mid y = A) = \mathcal{N}(x \mid \mu_A, 1)$
- 50% of the datapoints belong to class B and are distributed as $p(x \mid y = B) = \mathcal{N}(x \mid \mu_B, 1)$

What the maximum accuracy any classifier could achieve for this problem, depending on $\delta = \mu_A - \mu_B$? (you can assume $\mu_A > \mu_B$). The minimal possible error is also known as the *irreducible error* or *Bayes error rate.*

D [2p] Consider fitting a model on a new dataset. If we observe a very high training loss value, what does this tell us about the quality of the model? Is it over- or underfitting?

2 Bayesian Information Criterion (4 points)

The is commonly assumed in regression problems that the target variables y are generated by a deterministic function f and an additive, white noise error term ϵ, i.e.

$$y_i = f(x_i) + \epsilon_i \quad \text{where} \quad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

The goal is to recover the function f. Towards this goal, a parametric function $\hat{y}(x; \beta)$ is chosen, and the model is learned by maximizing the conditional likelihood $p(y \mid x) = \mathcal{N}(y \mid \hat{y}(x; \beta), \sigma^2)$.

A [2p] Show that for such a model, the conditional log-likelihood has the form

$$\ell(\beta, \sigma^2) = -\tfrac{1}{2\sigma^2}\,\mathrm{RSS} - \tfrac{1}{2} N \log\big(2\pi\sigma^2\big)$$

B [2p] Show that the maximum likelihood estimate for σ^2 is $\hat{\sigma}^2 = \mathrm{MSE}(\hat{y}) = \frac{1}{N}\|y - \hat{y}(x; \beta)\|_2^2$.

3 Multi-output Linear Regression (8 points)

When we have multiple independent outputs in linear regression, the model is defined as

$$p(y \mid x, W) = \prod_{j=1}^{M} \mathcal{N}(y_i \mid w_j^T x_i, \sigma^2).$$

Since the likelihood factorizes across dimensions, so does the maximum likelihood estimator (MLE). Thus

$$\hat{W} = [\hat{w}_1, \ldots, \hat{w}_M]$$

where $\hat{w}_j = \left(X^T X\right)^{-1} Y_{:,j}$. In this exercise we apply this result to a model with a 2-dimensional response vector $y_i \in \mathbb{R}^2$.
Suppose we have the following binary data, $x_i \in \{0, 1\}$, and the following training data:

x	y
0	$(-1,-1)^T$
0	$(-1,-2)^T$
0	$(-2,-1)^T$
1	$(1,1)^T$
1	$(1,2)^T$
1	$(2,1)^T$

A [1p] Write down the model closed form.(Hint: You can define an embedding function ϕ for x such that $\phi(0) = (1,0)^T$ and $\phi(1) = (0,1)^T$)

B [7p] Solve for $\hat{W}$.