Machine learning for data science I

11 June 2025

Surname, name (all caps) _____

Student ID: _____

This is a closed book exam.

Write clearly and justify your answers.

Time limit: 90 min.

Question:	1	2	3	4	Total
Points:	25	25	25	25	100
Score:					

1. Answer the following questions about Quantile Regression.

[7] (a) Linear regression minimizes the sum of squared residuals, which corresponds to estimating the mean of the distribution $p(y|x)$. What does minimizing the sum of absolute differences $\sum |y_i - \hat{y}_i|$ correspond to? Explain your answer.

[8] (b) What would be the effect of using Huber loss (with parameter δ) $L_\delta(y_i - \hat{y})$ compared to the squared and absolute loss mentioned previously?

$$L_\delta(a) = \begin{cases} \frac{1}{2}a^2 & \text{for } |a| \leq \delta, \\ \delta \cdot \left(|a| - \frac{1}{2}\delta\right) & \text{otherwise.} \end{cases}$$

[10] (c) Generalize the absolute difference loss from the first question with an appropriate loss function that will estimate the $\tau \cdot 100$-th percentile of the distribution $p(y|x)$. Explain your reasoning.

Hint: consider assigning different weights to observed values that are smaller/larger that the linear estimate $\hat{y}$.

2. Consider a convolutional neural network, which accepts a 4x4 image (single channel) X as the input, applies a convolutional layer with a 2x2 kernel K (with no bias, 0 padding and a stride of 1), uses a max pooling layer (again with 0 padding and a stride of 1) and finishes with a fully connected layer F (with bias b) to obtain a single regression output o. There are no activation functions used in the network.

[10] (a) The parameters are initialized to the following values:
$K = [[2, 1], [1, 0]]$, $F = [[1, 2], [-3, 5]]$, $b = -1$.
What is the output o of the neural network for
$X = [[1, 0, -1, 1], [0, 1, 1, -1], [2, 1, 0, 0], [1, 0, 1, 0]]$?

[15] (b) We will use MSE as a loss function. Suppose we make one step of gradient descent training (using a 0.5 learning rate) with the previous data instance X that has a correct output $y = 15$. What is the new value of $K_{0,0}$ (top-left weight of the convolution kernel) after this training step? Explain your calculation.

Hint: Don't compute the updates of the entire network but just enough to determine the change of $K_{0,0}$.

3. Answer the following questions about Shapley values.

[5] (a) Define the Shapley value of a feature in the context of explaining the contribution of features of a predictive model for a particular observation of feature values.

[5] (b) State and briefly explain the 4 axioms that uniquely define the Shapley value.

[10] (c) We have a simple data generating process with 3 binary features A, B, and C. The three features are independent, but A is 1 with $\frac{2}{3}$ probability, B is 1 with $\frac{1}{3}$ probability, and C is 1 with $\frac{8}{9}$ probability. Our simple classifier is $f(A = a, B = b, C = c) = a \vee b$ (the prediction is 1 if at least one of the features A and B is 1). Compute the Shapley values for A, B, and C for the instance $A = 0$, $B = 0$, $C = 0$.

[5] (d) In the previous subproblem you can compute expected predictions without knowing some of the features exactly, because you know the underlying distribution of features (also a hint for problem (c), so that you don't go off track). In practice, however, you typically have a sample dataset, but you do not know the exact distribution. How can we estimate expected predictions with missing features in practice?

4. Answer the following questions about kernel methods.

[10] (a) Explain the kernel trick in your own words. Why is it referred to as a "trick"? What problem does it solve?

[15] (b) In kernelized logistic regression, assume the weight vector can be expressed as a linear combination of the training examples. Show that the class probability prediction can be written as:

$$P(y = 1 \mid x) = \sigma\left(\sum_{i=1}^{n} \alpha_i K(x_i, x) + b\right)$$

where $\sigma(z)$ is the standard logistic function, and $K(x_i, x)$ is a kernel function.