Machine learning for data science I

17 June 2022

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	5	Total
Points:	20	20	20	20	20	100
Score:						

[20] 1. We have a classification problem with n examples from three equally represented classes. We have decided to use a simple classification model, which ignores the features and simply computes the distribution of class values in the training data. We want to evaluate it by observing its log loss and classification accuracy (CA). For measuring CA, the model predicts the most likely class value. Estimate the expected log loss and CA scores obtained with leave-out-out and 3-fold cross-validation. What are these scores when the number of examples is very small ($n = 3$) and very large? Explain your answers.

   1. leave-one-out
      (a) log loss
      (b) CA
   2. 3-fold cross-validation
      (a) log loss
      (b) CA

[20] 2. We have a regression data set with a single independent variable.

x_i	y_i
-2	6
-1	4
1	4
2	7
3	13

1. Roughly estimate the coefficients and intercept of a linear regression model. Report the mean squared error of your model on the given data.
2. Construct a new feature that will significantly improve your model. Roughly estimate and report the new model and its mean squared error.

[20] 3. Answer the following questions about string kernels.

1. Show that a function that counts the number of common bigrams in two words is a valid kernel function.

```
def bigrams(s):
    return set(s[i:i+2] for i in range(len(s)-1))

def kernel(s,t):
    return len(bigrams(s).intersection(bigrams(t)))
```

2. Show that a function that counts the number of common words in two sentences is a valid kernel function. Note that a word can repeat several times in both sentences.

```
def words(s):
    return s.split()

def kernel2(s,t):
    return sum(a==b for a in words(s) for b in words(t))
```

[20] 4. Consider a 1D problem with points in two clusters $A = \{A_1, \ldots, A_{n_A}\}$ and $B = \{B_1, \ldots, B_{n_B}\}$. We will compare the *average linkage* between clusters A and B ($D(A, B)$) with the distance between the centers (means) of clusters A and B ($d(\bar{a}, \bar{b})$).

1. Prove that they are the same in case of Euclidean distance between points.
2. Show that they are *not* the same if we were dealing with a 2D problem and use Euclidean distances.
3. What is the relation between D and d in terms of variance if we use a squared Euclidean distance $d(A_i, B_i) = (A_i - B_i)^2$?

[20] 5. We are interested in estimating parameter θ and are given a posterior distribution $p(\theta)$ and the following loss function:

$$L(\hat{\theta}, \theta) = \begin{cases} c_1|\hat{\theta} - \theta| & \text{if } \hat{\theta} \leq \theta \\ c_2|\hat{\theta} - \theta| & \text{if } \hat{\theta} > \theta \end{cases}$$

where $c_1, c_2 > 0$ are known constants.

Derive the estimator $\hat{\theta}$ of θ that minimizes Bayesian risk for this loss function.

Hint: you might find the Leibniz integral rule useful. In this case it is valid even for an infinite boundary (e.g. $a(x) = -\infty$) if we consider its derivative to be 0 ($\frac{d}{dx}a(x) = 0$).

$$\frac{d}{dx}\left(\int_{a(x)}^{b(x)} f(x,t)\,dt\right) = f\big(x, b(x)\big) \cdot \frac{d}{dx}b(x) - f\big(x, a(x)\big) \cdot \frac{d}{dx}a(x) + \int_{a(x)}^{b(x)} \frac{\partial}{\partial x} f(x,t)\,dt$$