Machine learning for data science I

12 June 2023

Surname, name (all caps) ____

Student ID: ____

This is a closed book exam.

Write clearly and justify your answers.

Time limit: 105 min.

Question:	1	2	3	4	5	Total
Points:	20	20	20	20	20	100
Score:						

1. Consider the absolute loss function for the discrete case $l(p, y) = \sum_{i=1}^{C} |p_i - y_i|$. C is the number of class values, $p = [p_1, \ldots, p_C]$ is the predicted probability distribution of class values and $y = [y_i, \ldots, y_C]$ is the outcome, where all y_i except one are zero. Use $r = [r_1, \ldots, r_C]$ as the true distribution of the data generative process.

[10] (a) This is an improper scoring rule. Explain what this means and describe other types of scoring rules.

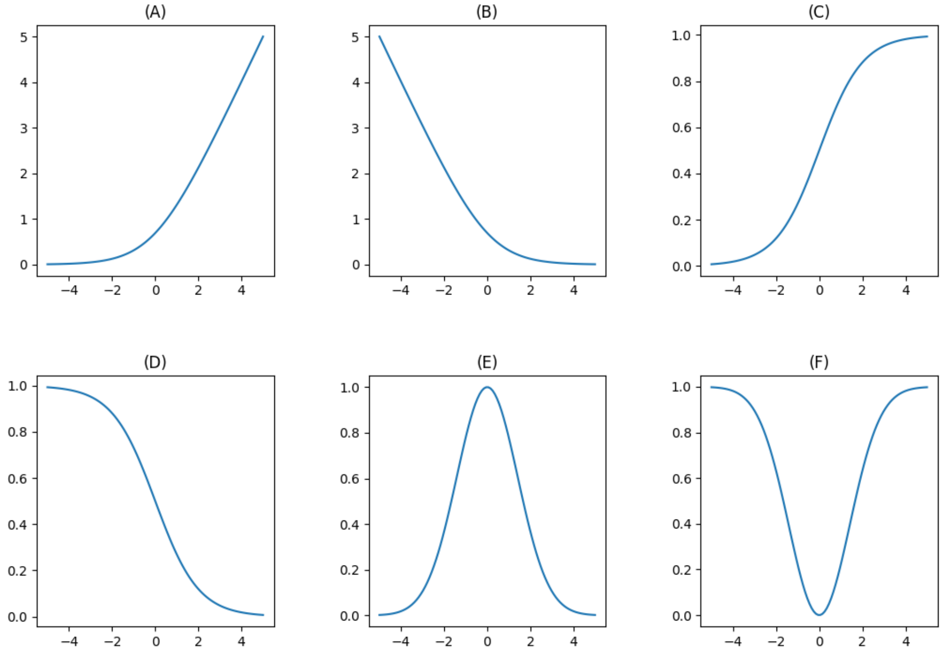
[10] (b) Which distribution minimizes the risk with this loss function? Take into account multiple class values C (not just 2). Prove your answer.

2. Consider a family of classification models with decision function $H(x) = \text{sign}(F(x))$, $F : \mathbb{R}^d \to \mathbb{R}$, $H : \mathbb{R}^d \to \{-1, 1\}$. We will optimize the parameters β of function F by minimizing the expression $\sum_i L(y_i \cdot F(x_i))$. Explain your answers to the following questions.

[6] (a) Which of the functions illustrated below can be used as a loss function L for the described classifier?

[4] (b) Which loss function is the most rebust to outliers?

[10] (c) Logistic regression is part of this familiy. What are the functions L and F which result in logistic regression? Which of the graphs corresponds to this loss function L?
Help: $p(y = 1|\beta) = \sigma(z)$, $\quad p(y = -1|\beta) = \sigma(-z)$.



3. Derive an upper bound of the error rate (ϵ_{NN}) of a 1-nearest neighbor classifier in terms of the error rate (ϵ_{OPT}) of a Bayes optimal classifier as the number of instances approaches infinity ($n \to \infty$). The class variable has $m > 1$ possible values.

[4] (a) What is the prediction of a Bayes optimal classifier? What is its error rate?

[8] (b) Show that $\epsilon_{NN} \leq 2\epsilon_{OPT}$.

[8] (c) Prove a tighter bound $\epsilon_{NN} \leq \epsilon_{OPT}(2 - \frac{m}{m-1}\epsilon_{OPT})$.

Help: you might find the following form of the Cauchy–Schwarz inequality useful.

$$(\sum_{i=1}^{m} p_i)^2 \leq m \sum_{i=1}^{m} p_i^2$$

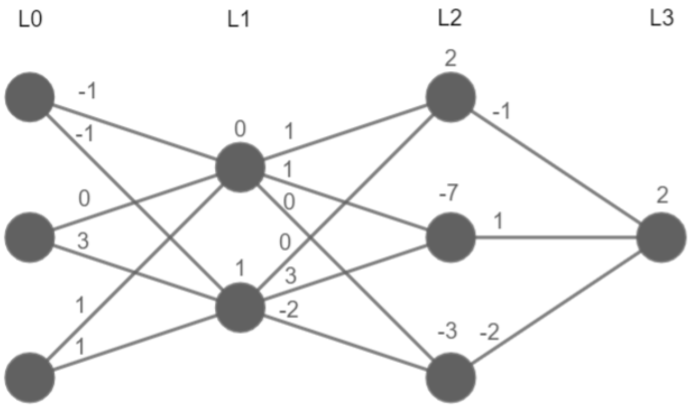
a) The most frequent class.

4. Given is a feed-forward neural network with two hidden layers ($L1$ and $L2$), input layer ($L0$), and output layer $L3$ (see figure). Let us denote the weighted sum of the inputs at layer i with vector $z^{[i]}$ and activations with vector $a^{[i]}$. Let $W^{[i]}_{jk}$ denote the weight connecting j-th unit of layer $i-1$ and k-th unit of layer i (illustrated above the lines). $B^{[i]}_j$ denotes the bias of the j-th unit in layer i (illustrated above the nodes). All units use $ReLU$ activation functions except for the output layer, which doesn't use an activation function.

[6] (a) What is the output $\hat{y}$ of the neural network for a data instance $x = [2, 2, -1]$?

[14] (b) We aim to minimize the loss function $J = \frac{1}{2}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2$, where m indicates the number of data instances. Suppose we make one step of training (with learning rate 0.5) the illustrated network with a single data instance $x = [2, 2, -1]$ that has a correct output $y = 2$. What is the new value of $W^{[1]}_{0,1}$ (weight between the top unit of level 0 and the bottom unit of level 1) after this training step? Explain your calculation.

Hint: Think about which parts of the computation you can skip but don't forget to explain why.



5. Bayesian statistics basics. Some help with distributions:

Bernoulli (pmf): $P(x = k) = \begin{cases} 1 - p & \text{if } k = 0 \\ p & \text{if } k = 1 \end{cases}$, $E[x] = p$

Beta (pdf): $P(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$, $E[x] = \frac{\alpha}{\alpha+\beta}$

[4] (a) Write Bayes' Theorem for the posterior $p(\theta|y)$, given likelihood $p(y|\theta)$ and prior $p(\theta)$.

[5] (b) Derive the posterior for θ given n observations $y = [y_1, \ldots, y_n]$ if the likelihood is Bernoulli(θ) and the prior is Beta(α, β).

[5] (c) Derive the definition of the posterior predictive $p(y'|y)$ of a new observation y' given a set of n observations y. Express it with the likelihood and posterior only, explain the steps.

[6] (d) Derive the posterior predictive for the Bernoulli-Beta case from above.