Machine learning for data science I

22 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 a linear regression model with a Laplace prior on the parameters. Denote the number of data samples with n and the number of features with k.

   Help: probability density function of Laplace distribution is $f(x|\mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right), b > 0$

[6] (a) What is the purpose of regularization in machine learning models? What kind of regularization does a Laplace prior in linear regession correspond to? Write the cost function that we want to minimize in such regularized linear regression.

[4] (b) What is the approximate relation between the distribution parameters (μ, b) and regularization weight (λ) - which parameters are proportional, inversely proportional or neither and why?

[10] (c) Derive an exact relation between μ, b and λ and prove your answers to the previous questions in the process.

a) TO REDUCE OVERFITTING

TO INFER SIMPLER MODELS

L2 REGULARIZATION

$$\sum_{i=1}^{n} (y_i - \beta^T x_i)^2 + \lambda \sum_{i=1}^{p} ||\beta_i||$$

b) $\mu = 0$ (WE ASSUME THAT IT IS)

SMALL b => NARROW DISTR. => STRONG REG => LARGE λ

=> INVERSE PROP.

c) $$\beta^* = \arg\max_\beta \; p(y|\beta) \cdot p(\beta)$$

$$= \underbrace{\prod p(y_i|\beta)}_{\text{NORMAL}} \cdot \prod p(\beta_j)$$

$$= \prod \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{(y_i - \beta^T x_i)^2}{2\sigma^2}\right)} \cdot \prod \frac{1}{2b} e^{-\frac{\beta_i}{b}}$$

$-$ LOG

$$= \arg\min_\beta \sum \frac{(y_i - \beta^T x_i)^2}{2\sigma^2} + \sum \frac{|\beta_i|}{b} = \arg\min_\beta \sum (y_i - \beta^T x_i)^2 + \underbrace{\frac{2\sigma^2}{b}}_{\lambda} \sum |\beta_j|$$

2. Bootstrap aggregating.

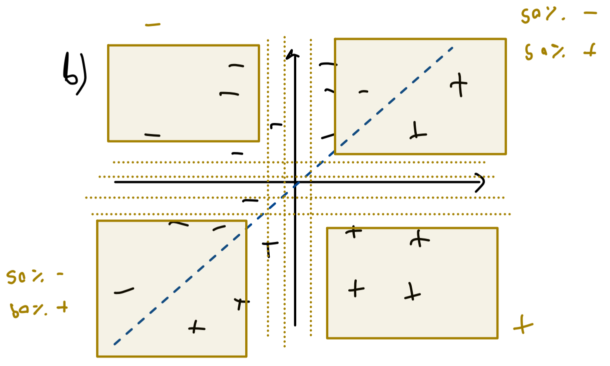
[6] (a) Explain what is bagging (Bootstrap aggregating). Describe two advantages and two disadvantages of bagging.

[7] (b) Consider a 2D classification problem $y = x_1 > x_2\,(x_1, x_2 \in \mathbb{R},\, y \in \{0, 1\})$ with $n = 1000$ data points. How does bagging (with $m = 100$ datasets) using classification trees of depth 1 (single split) perform on such data and why?

[7] (c) Compute the expected number of distinct instances in the bootstrap sample as the ratio of the original data set with n cases as $n \to \infty$. We are interested in an exact solution (we know that it is approximately 60%).

Help: $\lim_{n\to\infty}(1 + x/n)^n = e^x$. Consider using indicator variables I_i, which have a value 1 if the i-th instance is included in the bag and 0 otherwise.

b)



SPLITS WILL BE CLOSED TO 0 =>

c)

$I_i = 0/1$

$x = \sum I_i$

$\lim_{x\to\infty} \frac{E[x]}{n} \Rightarrow E[x] = E[\sum I_i] = \sum E[I_i] = n \cdot E[I_i]$

$E[I_i] = p(I_i = 1) = 1 - p(I_i = 0)$

$p(I_i = 0) = \left(\frac{n-1}{n}\right)^n = \left(1 - \frac{1}{n}\right)^n$

$\Rightarrow \lim_{x\to\infty} \frac{\cancel{n} E[I_i]}{\cancel{n}} = \lim_{x\to\infty} 1 - \left(1 - \frac{1}{n}\right)^n = 1 - \lim_{x\to\infty}\left(1 - \frac{1}{n}\right)^n = 1 - \frac{1}{e}$

3. Artificial neural networks (ANN) with ReLU activation function are universal piecewise-linear approximators.

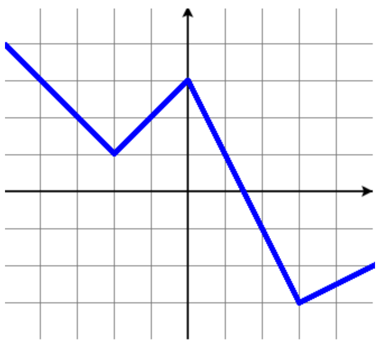
[8] (a) What is a ReLU activation function? Why does ANN need activation functions? Which other activation functions do you know (describe two other besides linear and ReLU)?

[6] (b) Consider only the two linear segments on the left of the illustration in the image below, that is a function

$$f(x) = \begin{cases} -x - 1 & x < -2 \\ x + 3 & x >= -2 \end{cases}$$

defined on $x \in \mathbb{R}$. Define the neural network (architecture and parameters) that corresponds to such piecewise-linear approximation. Describe your construction process.

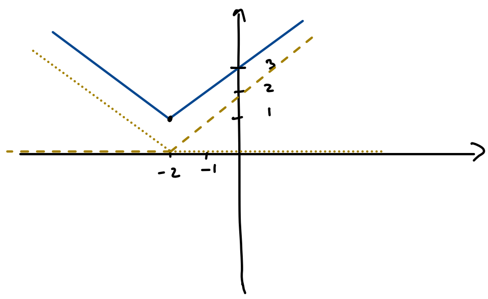
[6] (c) Extend your piecewise-linear approximaton to the entire function in the illustration defined on $x \in \mathbb{R}$.



a) $ReLU(x) = max(x, 0)$

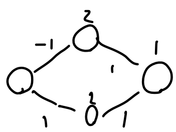
SIGMOID, TANH ...

b)

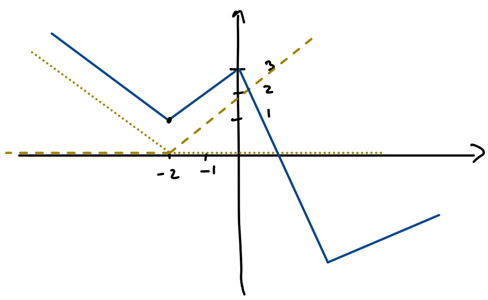


$f(x) = max(-x-2, 0) + max(x+2, 0) + 1$

NETWORK



c)



$f(x) = max(-x-2, 0) + max(x+2, 0) + 1 - max(3x, 0)$
$+ max(2.5x - 7.5, 0)$

4. Answer the following questions about kernels.

[6] (a) What does a Support Vector Machine optimize? What is the purpose of using kernels with SVM? List two often used kernels in SVM.

[7] (b) Consider a kernel $K(S,T) = e^{|S \cap T|}$ defined on two sets $S, T \subseteq U$. Prove that it is a valid (Mercer) kernel.

[7] (c) Cosine similarity between two documents A and B is defined as a cosine of the angle θ between the corresponding vectors a and b representing the number of occurrences of each word in the document. Prove that cosine similarity is a valid kernel function.

b)

$$
\begin{array}{cc}
 & \begin{array}{ccccc} U_1 & U_2 & U_3 & U_4 & U_5 \end{array} \\
S = & \begin{bmatrix} 0 & 1 & 1 & 0 & 1 \end{bmatrix}
\end{array}
= \{U_2, U_3, U_5\}
$$

$$T = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 \end{bmatrix} = \{U_1, U_2, U_4, U_5\}$$

$$S \cdot T = 0 + 1 + 0 + 0 + 1 = 2$$

$\Rightarrow$ DOT (S,T) GIVES THE SIZE OF THE INTERCEPTION

↳ IT IS A VALID KERNEL

c)

$$k(A,B) = \cos\varphi = \frac{A \cdot B}{\|A\| \cdot \|B\|} = \underbrace{f(A) \cdot f(B) \cdot A \cdot B}_{k'(A,B) \text{ is valid}} = \frac{A}{\|A\|} \cdot \frac{B}{\|B\|}$$

$\Rightarrow$ $k = f(A) \cdot f(B) \cdot k'$ is also valid

5. Principal Component Analysis

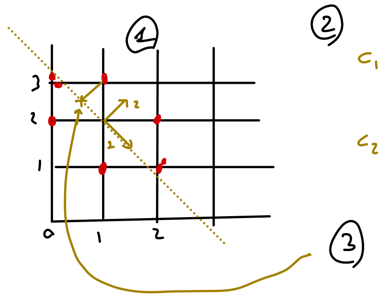
[11] (a) Consider a data set X consisting of points $[(0,2),(0,3),(0,3),(1,1),(1,3),(2,1),(2,1),(2,2)]$. Note that some of them appear more than once.

1. Provide a short description of the PCA technique.
2. Determine the principal components for the given dataset - you can do so visually, but explain your process.
3. Compute the PCA approximation of the point $(1,3)$ using only the first principal component.

[9] (b) We have a large data set $X = [x_1, \ldots, x_n]$, $x_i \in \mathbb{R}^d$ and are considering making some modifications to it.

1. We will standardize every dimension (to zero mean and unit variance) before doing PCA. Describe one situation where this makes sense and one where it doesn't.
2. We will introduce another dimension that will be equal to 1 for all data points. How does this change affect the principal components?
3. We discovered that there are some binary labels assigned to the data points. We want to perform dimensionality reduction to two dimension with PCA and then train a classification model on such 2D data set. Which principal components do we select?

a)



2

$c_1 = \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right)$

$c_2 = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$

b)