Calculators may be used in this examination provided they are not capable of being used to store alphabetical information other than hexadecimal numbers

UNIVERSITY OF BIRMINGHAM

School of Computer Science

Machine Learning

Main Summer Examinations 2024

Time allowed: 2 hours

[Answer all questions]

Turn Over

Note

Answer ALL questions. Each question will be marked out of 20. The paper will be marked out of 60, which will be rescaled to a mark out of 100.

The end of the paper has an appendix with some formulas and definitions that you may find useful.

Question 1 Core Concepts

(a) A machine learning algorithm can be seen as an algorithm that takes a training set and a hypothesis set as inputs.

- What is a hypothesis set?
- Give an example of hypothesis set. You must include a mathematical definition and an explanation of your example. Your example may correspond to an existing machine learning approach or to a fictitious approach, so long as the hypothesis set is appropriately defined and explained.

[5 marks]

(b) Why can it be useful to adopt the dual representation of the support vector machine problem instead of the primal representation? Explain your answer in detail.

[5 marks]

(c) For a given regression problem, you train a model on a training set and achieve a loss of zero. Then, you apply the model to test examples and find that it is making mistakes in its predictions.

- Explain the meaning of bias and variance in this context.
- Then, discuss a possible reason for your observation, using bias and variance.

[5 marks]

(d) L_2 regularisation is sometimes used to control the bias and variance of the solution to regression problems. It is often referred to as a shrinkage method.

- What is meant by the term "shrinkage method"?
- Give an example of another shrinkage method and explain how it influences the solutions of regression problems.

[5 marks]

Question 2 Classification

(a) Assume that you were testing an implementation of hard margin support vector machines based on sequential minimal optimisation, without any feature transformation. You adopted the training set below, which is composed of linearly separable examples:

$\mathbf{x}^{(1)} = (1, 3)^T$, $y^{(1)} = +1$ $\quad$ $\mathbf{x}^{(4)} = (2, 1)^T$, $y^{(4)} = -1$

$\mathbf{x}^{(2)} = (2, 4)^T$, $y^{(2)} = +1$ $\quad$ $\mathbf{x}^{(5)} = (3, 2)^T$, $y^{(5)} = -1$

$\mathbf{x}^{(3)} = (0.5, 3.5)^T$, $y^{(3)} = +1$ $\quad$ $\mathbf{x}^{(6)} = (4, 0.5)^T$, $y^{(6)} = -1$

The implementation retrieved the decision boundary $h(\mathbf{x}) = x_2 - x_1 = 0$, with $(\mathbf{x}^{(1)}, y^{(1)})$ and $(\mathbf{x}^{(4)}, y^{(4)})$ as support vectors. Even though all training examples are correctly classified, you suspect that the implementation may have some bug as it was unable to successfully solve the underlying optimisation problem. Mathematically show that the optimisation problem was not successfully solved. Explain your reasoning. **[10 marks]**

(b) Consider a binary classification problem with two input variables and whose true decision boundary is $f(\mathbf{x}) = e^{x_1} + 3x_2^2 + 1 = 0$. You wish to learn a binary classification model for this problem using logistic regression with feature transformation.

- Propose a suitable feature transformation to use.
- Prove that your proposed feature transformation is suitable. Explain your proof.

[10 marks]

Question 3 Learning Theory

In learning problems, we have a set of training samples denoted by $\mathcal{D}$ and all samples are independent and identically distributed (i.i.d.) random variables taken from an unknown distribution. Learning is to find an unknown target function f. Suppose that we want to explore an entire hypothesis set $\mathcal{H}$, looking for some $h \in \mathcal{H}$ that has a small error. After learning, we denote the final hypothesis as g.

(a) Assume we have two hypotheses h_1 and h_2 for a regression problem, where h_2 is more complex than h_1. Explain the meaning of the approximation-generalisation trade-off using these two hypotheses in the current context. Illustrate your answer with appropriate diagrams. **[10 marks]**

(b) Consider the Hoeffding Inequality

$$\mathbb{P}\left(|E_{in}(g) - E_{out}(g)| > \epsilon\right) \leq 2Me^{-2\epsilon^2 N}.$$

(i) Explain the meaning of each term, $E_{in}(g)$, $E_{out}(g)$, ϵ, M, N, in this expression. Then explain the meaning of this expression. **[6 marks]**

(ii) The VC analysis tries to replace M by another term involving d_{VC}. Explain the reason for this manipulation. **[4 marks]**

Appendix

Dual representation of hard margin support vector machines

$$\underset{\text{a}}{\text{argmax}} \tilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a^{(n)} - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a^{(n)} a^{(m)} y^{(n)} y^{(m)} k(\mathbf{x}^{(n)}, \mathbf{x}^{(m)})$$

$$\text{subject to} \quad a^{(n)} \geq 0, \ \forall n \in \{1, \cdots, N\} \text{ and } \sum_{n=1}^{N} a^{(n)} y^{(n)} = 0,$$

where $a^{(i)}$ is the Lagrange multiplier associated to training example i, N is the number of training examples, $\mathbf{x}^{(i)} \in R^d$ are the input variables of example i, d is the number of input variables, $y^{(i)} \in \{-1, 1\}$ is the output label of example i, and $k(\cdot, \cdot)$ is the kernel function.

Primal representation of hard margin support vector machines

$$\underset{\text{w},b}{\text{argmin}} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 \right\}$$

$$\text{subject to} \quad y^{(n)} h(\mathbf{x}^{(n)}) \geq 1, \quad \forall (\mathbf{x}^{(n)}, y^{(n)}) \in \mathcal{T},$$

where $\mathbf{w}$ and b are the parameters to be learned, $\mathbf{x}^{(i)} \in R^d$ are the input variables of example i, d is the number of input variables, $y^{(i)} \in \{-1, 1\}$ is the output label of example i, $\mathcal{T}$ is the training set, $h(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b$, and $\phi(\mathbf{x})$ is a feature embedding. When $\phi(\mathbf{x}) = \mathbf{x}$, no embedding is being used.

Perpendicular distance between a point and a hyperplane

The perpendicular distance between a point and a hyperplane $h(\mathbf{x}) = 0$ representing a correct decision boundary for the hard margin support vector machines can be defined as:

$$\text{dist}(h, \mathbf{x}^{(n)}) = \frac{|h(\mathbf{x}^{(n)})|}{\|\mathbf{w}\|} = \frac{y^{(n)} h(\mathbf{x}^{(n)})}{\|\mathbf{w}\|}.$$

Do not complete the attendance slip, fill in the front of the answer book or turn over the question paper until you are told to do so

Important Reminders

- Coats/outwear should be placed in the designated area.
- Unauthorised materials (e.g. notes or Tippex) must be placed in the designated area.
- Check that you do not have any unauthorised materials with you (e.g. in your pockets, pencil case).
- Mobile phones and smart watches **must** be switched off and placed in the designated area or under your desk. They must not be left on your person or in your pockets.
- You are not permitted to use a mobile phone as a clock. If you have difficulty seeing a clock, please alert an Invigilator.
- You are not permitted to have writing on your hand, arm or other body part.
- Check that you do not have writing on your hand, arm or other body part – if you do, you must inform an Invigilator immediately
- Alert an Invigilator immediately if you find any unauthorised item upon you during the examination.

Any students found with non-permitted items upon their person during the examination, or who fail to comply with Examination rules may be subject to Student Conduct procedures.