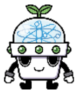
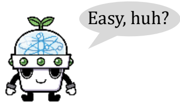


Statistical Learning Theory
Final Exam 2022



** READ THE FOLLOWING INSTRUCTIONS CAREFULLY **

(There is a risk that your answer will not be graded correctly if the instructions are not followed)

- The exam has two parts (PART I and PART II)
- Use the first answer sheet for PART I, and the second sheet for PART II.
- You can use both sides of each sheet.
- Write your name and ID on the both answer sheets.
- Answer all of the questions in English.

PART I

Q.1 Fill in the blanks.

(1) Ridge regression is L[ ]-regularized linear regression

(2) Ridge regression can be interpreted as [ ] estimation a Bayesian inference framework under some assumptions.

(3) [ ] is used as a measure of the complexity of a classifier class of infinite size.

(4) One example of a real-world applications of the multi-class classification problem is [ ].

(5) The [ ] loss is a convex upper bound of the zero-one loss.

Q.2 Let us consider a pairwise comparison problem. We have n training data instances $\left\{\left(\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}, y_i\right)\right\}_{i=1,2,\ldots,n}$, where, for each $i \in \{1,2,\ldots,n\}$, $\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)} \in \mathbb{R}^D$ denote the feature vectors of two input objects sampled in an i.i.d manner. $y_i \in \{+1,-1\}$ indicates which of the two objects is ranked higher than the other; namely, $y_i = +1$ indicates $\mathbf{x}_i^{(1)}$ is superior to $\mathbf{x}_i^{(2)}$, and $y_i = -1$ indicates the opposite. We consider the following model that gives the conditional probability $p(y = +1|\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$ of the comparison label y being $+1$ given inputs $\mathbf{x}^{(1)}, \mathbf{x}^{(2)} \in \mathbb{R}^D$, which is defined as

$$p(y = +1|\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = \frac{\exp(\mathbf{w}^\top \mathbf{x}^{(1)})}{\exp(\mathbf{w}^\top \mathbf{x}^{(1)}) + \exp(\mathbf{w}^\top \mathbf{x}^{(2)})},$$

where $\mathbf{w} \in \mathbb{R}^D$ is the model parameter and $\top$ indicates the transpose of a vector .

(1) Give the objective function (to maximize) for estimating $\mathbf{w}$ by maximum likelihood estimation.

(2) Give a stochastic gradient descent update formula (i.e., steepest gradient descent using only $\left(\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}, y_i\right)$) for the objective function you gave in Q.1. (Note that this is actually gradient “ascent” because the Q.1 is a maximization problem.)

(3) Now we consider replacing the above model using a neural network, that is, $p\left(y = +1\middle|\mathbf{x}^{(1)}, \mathbf{x}^{(2)}\right) = f(\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$. What is a potential concern in such modeling? And, what is a possible way to address this issue?

PART II

Q.3 We have n data instances $\{\mathbf{x}_i\}_{i=1,2,\ldots,n}$, where for each $i \in \{1,2,\ldots,n\}$, $\mathbf{x}_i \in \mathbb{R}^D$. Assume that $\sum_{i=1}^{n} \mathbf{x}_i = \mathbf{0}$. Show that the principal component analysis (PCA) is equivalent to linear autoencoder model for an orthonormal matrix.

$$\widehat{\boldsymbol{U}} = \operatorname{argmax} \sum_{i=1}^{n} tr(\boldsymbol{U}^{\boldsymbol{T}} \mathbf{x}_i \, \mathbf{x}_i^T \boldsymbol{U}), s.t. \boldsymbol{U}^{\boldsymbol{T}} \boldsymbol{U} = \boldsymbol{I}$$

Q.4 Let us denote $p'(\mathbf{x}, y)$ a probability density function with $p(\mathbf{x}, y) \neq p'(\mathbf{x}, y)$. Derive the empirical risk of J' under the assumption $p(y|\mathbf{x}) = p'(y|\mathbf{x})$ using $\{(\mathbf{x}_i, y_i)\}_{i=1,2,\ldots,n} \sim p(\mathbf{x}, y)$ and the importance weight function $r(\mathbf{x}) = p'(\mathbf{x})/p(\mathbf{x})$.

$$J' = -\iint (y \log(f(\mathbf{x})) + (1-y)\log(1-f(\mathbf{x}))) \quad p'(\mathbf{x}, y) \mathrm{d}\mathbf{x} \, \mathrm{d}y$$

Q.5 Explain the key difference between the wrapper method and the filter method in feature selection.

Q.6 Explain how to formulate the node classification problem using a Graph Neural Network model with equations.