CMPSC 448: Machine Learning & Algorithmic AI

Mock Exam
Spring 2023

Question 1 (Multiple Choice Questions) *Fill in the bubble for the correct answer (there might be multiple correct answers).*

a) Which of the following algorithm(s) is guaranteed to return a unique solution:

◯ Hard Support Vector Machines

◯ Logistic Regression

◯ Ridge Regression

◯ Lasso Regression

b) Underfitting can happen if we try to learn a complex unknown model with a simple hypothesis class.

◯ `True`

◯ `False`

c) We consider a classification problem on linearly separable data. Our dataset had an outlier-a point that is very far from the other datapoints in distance (and also far from margins in SVM but still correctly classified by the SVM classifier).

We trained the SVM, logistic regression and 1-nearest-neighbour models on this dataset. We tested trained models on a test set that comes from the same distribution as training set, but doesn't have any outlier points. After that we removed the outlier and retrained our models.

After retraining, which classifier will change its decision boundary around the test points.

◯ Logistic regression

◯ SVM

◯ 1-nearest-neighbors classifier

◯ All of them

d) Consider the k-fold cross validation on a linear regression model with a sufficiently large amount of training data. When k is large, the computational complexity of the k-fold cross validation with respect to k is of order

◯ $O(1/K)$

◯ $O(K)$

◯ $O(1)$

◯ $O(K(K-1))$

e) How does the bias-variance decomposition of a ridge regression estimator compare with that of the ordinary least-squares estimator in general?

◯ Ridge has a smaller bias, and smaller variance.

◯ Ridge has a larger bias, and larger variance.

◯ Ridge has a larger bias, and smaller variance.

◯ Ridge has a smaller bias, and larger variance.

f) Which of the following statements is true about nearest neighbor classifiers:

◯ Nearest neighbors can be slow to find in high-dimensional spaces.

◯ Nearest neighbor classifiers can only work with the Euclidean distance.

◯ Nearest neighbor classifiers do not need to store the training data.

Question 2 (Short Answer Questions) *Answer the following questions. Your answer should be brief and precise.*

a) Assume that we want to fit an affine line through a given point $(x_1, y_1) = (1, 1) \in \mathbb{R}^2$. To do so, we want to minimize the function $f(w_0, w_1) = (y_1 - (w_0 + w_1 x_1))^2$ using gradient descent from a starting point $[0, 0]$. Using an appropriate and strictly positive step-size, the iterates $[w_{0,t}, w_{1,t}]$, will converge to what solution.

b) Consider a linear regression problem with n samples where the input is in d-dimensional space, and all output values are $y_i \in \{-1, +1\}$. Which of the following statements is correct? Justify your answer.

- ◯ Linear regression cannot "work" if $n \gg d$
- ◯ Linear regression cannot "work" if $n \ll d$
- ◯ Linear regression can be made to work perfectly if the data is linearly separable

Question 3 (Weighted Regression)

In this problem, we generalize the Ridge Regression to incorporate the importance of individual training samples. Specifically, besides training set $S = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots, (\boldsymbol{x}_n, y_n)\}$ where $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$, we are also given positive weights $a_1, a_2, \dots, a_n$ where a_i is the importance (weight) of ith training data. Show that the solution to the weighted least squares problem with ℓ_2 regularization

$$\boldsymbol{w}_* = \arg\min_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{2}\sum_{i=1}^{n} a_i \left(\boldsymbol{w}^\top \boldsymbol{x}_i - y_i\right)^2 + \frac{\lambda}{2}\|\boldsymbol{w}\|^2$$

is given by the formula

$$\boldsymbol{w}^* = \left(\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X} + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{y}$$

where $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ is the data matrix (defined in class), $\boldsymbol{y} \in \mathbb{R}^n$ is the labels vector, $\boldsymbol{I}$ is $d \times d$ identity matrix, and $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is a diagonal matrix where $A_{ii} = a_i$, i.e.,

$$\boldsymbol{A} = \begin{bmatrix} a_1 & 0 & \dots & 0 \\ 0 & a_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & a_n \end{bmatrix}.$$

(Scratch Space)

(Scratch Space)