MCE5901 Homework Set 4

November 5, 2023

NOTICE: The homework is due on Nov. 17 (Friday) 11:59pm. **Please draft your solution with steps and provide the codes that you use for the applied problems.** You will need to submit two files to the blackboard: a PDF file including the solutions and results for all problems, and a file containing the codes for the applied problems. You are allowed, and even encouraged, to discuss the homeworks with your classmates. However, you must **write up the solutions on your own**. Plagiarism and other anti-scholarly behavior will be dealt with severely.

Problem 4 is optional. You will receive 20 points bonus if you solve it and the maximum score of the assignment is 100 points.

Problem 1. Maximal margin classifier

We explore the maximal margin classifier on a toy data set.

(a) We are given $n = 7$ observations in $p = 2$ dimensions. For each observation, there is an associated class label.

Obs.	X_1	X_2	Y
1	3	4	Red
2	2	2	Red
3	4	4	Red
4	1	4	Red
5	2	1	Blue
6	4	3	Blue
7	4	1	Blue

(b) Sketch the optimal separating hyperplane, and provide the equation for this hyperplane.

(c) Describe the classification rule for the maximal margin classifier. It should be something along the lines of "Classify to Red if $\beta_0+\beta_1 X_1+\beta_2 X_2 > 0$, and classify to Blue otherwise." Provide the values for β_0, β_1, and β_2.

(d) On your sketch, indicate the margin for the maximal margin hyperplane.

(e) Indicate the support vectors for the maximal margin classifier.

(f) Argue that a slight movement of the seventh observation would not affect the maximal margin hyperplane.

(g) Sketch a hyperplane that is not the optimal separating hyperplane, and provide the equation for this hyperplane.

(h) Draw an additional observation on the plot so that the two classes are no longer separable by a hyperplane.

Problem 2. Decision tree

The following table contains training examples that help predict whether a patient is likely to have a heart attack.

Patient ID	gender	chest pain	smoke	exercises	heart attack
1	male	yes	no	yes	yes
2	male	yes	yes	no	yes
3	female	no	yes	no	yes
4	male	no	no	yes	no
5	female	yes	yes	yes	yes
6	male	no	yes	yes	no

(a) Use classification error rate as criteria to construct a minimal decision tree (a single split) that predicts whether or not a patient is likely to have a heart attack. SHOW EACH STEP OF THE COMPUTATION.

(b) Use cross entropy as criteria to construct a minimal decision tree that predicts whether or not a patient is likely to have a heart attack. SHOW EACH STEP OF THE COMPUTATION.

Problem 3. Applied problem: SVM and classification tree

We will work on the Iris dataset and consider a classification problem to classify the types of iris. The Iris dataset is one of datasets scikit-learn comes with that do not require the downloading of any file from some external website. Use from sklearn.datasets import load_iris. Information about the dataset can be found from `https://scikit-learn.org/stable/auto_examples/datasets/plot_iris_dataset.html`.

(a) Split the data into train (75%) set and test (25%) set. Set random state = 42.

We first try support vector machine. We only take the first two features as input, i.e., Sepal length and Sepal width, and use these two features to classify the types of iris.

(b) Produce a scatter plot of the training data, ensuring that each class is labeled with a distinct color. Are the classes linearly separable?

(c) Set the SVM regularization parameter $C = 1$. Plot the decision surface for three SVM classifiers (using SVC module) with three different kernels: linear kernel, polynomial (degree 3) kernel and RBF kernel ($\gamma = 0.7$). How many support vectors for each model?

(d) Repeat (c) using $C = 5$. What do you find?

(e) Perform cross-validation using GridSearchCV() to select the best choice of C for SVM with linear kernel. For the best C, report the confusion matrix on the test data.

Next we try classification tree. Now we include all the features as input to classify the types of iris.

(f) Construct a single tree and plot it using Graphviz. What is the default classification criteria used?

(g) What is the most important feature? Translate the decision rules, i.e., export a textual format of the tree.

(h) Implement the bagging, random forest and Adaboost, show their performance in terms of error rate on the test sets. Comment on resutls.

Reference links: `https://scikit-learn.org/stable/modules/svm.html`
`https://www.science.smith.edu/~jcrouser/SDS293/labs/lab15-py.html`
`https://scikit-learn.org/stable/modules/tree.html`
`https://www.science.smith.edu/~jcrouser/SDS293/labs/lab14-py.html`

Problem 4. Kernels on discrete sequences (optional)

Thus far, we have dealt with kernel feature mappings that map real-valued inputs to some feature space. As we will see in this question, it is also possible to define feature mappings on discrete objects such as sequences of tokens. This is especially useful in applications such as bioinformatics and natural language processing.

Let $\mathcal{S}$ denote the set of strings (sequences) formed by concatenating elements of some alphabet $\mathcal{V}$. For example, $\mathcal{V}$ might be the set $\{a, b, \dots, z, \text{[space]}\}$ in the case of the English alphabet, in which case $\mathcal{S}$ contains all strings obtained by concatenating these characters (e.g. "the", "cat", "the cat", etc.). For any $s_1, s_2 \in \mathcal{S}$, we define

$$\delta(s_1, s_2) = \begin{cases} 1, & \text{if } s_1 = s_2 \text{ (i.e. the two strings are the same)} \\ 0, & \text{otherwise.} \end{cases}$$

(a) For any sequence $s \in \mathcal{S}$, let $\mathcal{G}_n(s)$ denote the set of all length n sub-sequences of s. For example,

$$\mathcal{G}_2(\text{abcd}) = \{\text{ab,bc,cd}\}.$$

We define the kernel function:

$$K_{\mathcal{G}}(s_1, s_2) = \sum_{x \in \mathcal{G}_n(s_1)} \sum_{z \in \mathcal{G}_n(s_2)} \delta(x, z)$$

Express the kernel function using an appropriate feature mapping $\phi(s)$. That is, define the feature mapping $\phi(s) \in \mathbb{R}^d$ such that $K_{\mathcal{G}}(s_1, s_2) = \phi(s_1)^T \phi(s_2)$. What is the dimension of the feature mapping space d? Hint: The description of $\phi(s)$ should be brief.

(b) Let $\mathcal{W}(s)$ denote the set of words in s (substrings in s separated by the [space] character). As an example,

$$\mathcal{W}(\text{machine learning is fascinating}) = \{\text{machine, learning, is, fascinating}\}.$$

We define a new kernel function $K_{\mathcal{W}}(s_1, s_2)$ as:

$$K_{\mathcal{W}}(s_1, s_2) = \sum_{x \in \mathcal{W}(s_1)} \sum_{z \in \mathcal{W}(s_2)} \delta(x, z).$$

Express this function using an appropriate feature mapping ϕ. That is, as in the previous part of this question, your task is to define the feature mapping $\phi(s)$ such that $K_{\mathcal{W}}(s_1, s_2) = \phi(s_1)^T \phi(s_2)$. What is the dimension of the feature mapping space in this case?