SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

50.007 Machine Learning, Summer 2023
Homework 3

Due 28 July 2023, 11:59 pm

1. Non-Linear SVM [10 Points]

Download and install the widely used SVM implementation LIBSVM (`https://github.com/cjlin1/libsvm` or `https://www.csie.ntu.edu.tw/~cjlin/libsvm/`; clicking on either link takes you to the webpage). We expect you to install the package on your own – this is part of learning how to use off-the-shelf machine learning software. Read the documentation to understand how to use it.

Download `promoters.zip`. In that folder are `training.txt` and `test.txt`, which respectively contain 74 training examples and 32 test examples in LIBSVM format. The goal is to predict whether a certain DNA sequence is a promoter[1] or not based on 57 attributes about the sequence (this is a binary classification task).

Run LIBSVM to classify promoters with different kernels (0-3), using default values for all other parameters. What is your test accuracy for each kernel choice?

2. Kernel Methods [16 Points]

In this problem, we consider constructing new kernels by combining existing kernels. Recall that for some function $K(\mathbf{x}, \mathbf{z})$ to be a kernel, we need to be able to write it as an inner product of vectors from some high-dimensional feature space:

$$K(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x})^T \phi(\mathbf{z})$$

Mercer's theorem gives a necessary and sufficient condition for a function K to be a kernel: its corresponding kernel matrix has to be symmetric and positive semidefinite, where the elements of a kernel matrix are inner products between all pairs of examples.
Suppose that $K_1(\mathbf{x}, \mathbf{z})$ and $K_2(\mathbf{x}, \mathbf{z})$ are kernels over $\mathcal{R}^n \text{x} \mathcal{R}^n$. For each of the cases below, state whether K is also a kernel. If it is, prove it. If it is not, give a counter example. (*Hints: You can use either Mercer's theorem or the definition of a kernel, as needed.*).

[1] A promoter is a region of DNA that facilitates the transcription of a particular gene. The ability to predict promoters is of practical importance in searching for new promoter sequences.

1. $K(\mathbf{x}, \mathbf{z}) = K_1(\mathbf{x}, \mathbf{z})K_2(\mathbf{x}, \mathbf{z})$
2. $K(\mathbf{x}, \mathbf{z}) = aK_1(\mathbf{x}, \mathbf{z}) + bK_2(\mathbf{x}, \mathbf{z})$, where $a, b > 0$ are real numbers
3. $K(\mathbf{x}, \mathbf{z}) = aK_1(\mathbf{x}, \mathbf{z}) - bK_2(\mathbf{x}, \mathbf{z})$, where $a, b > 0$ are real numbers
4. $K(\mathbf{x}, \mathbf{z}) = f(\mathbf{x})f(\mathbf{z})$, where $f : \mathcal{R}^n \to \mathcal{R}$ be any real valued function of x.

3. Logistic Regression [10 Points]

You are given a training set `diabetes_train.csv`. Each row in the file contains the probability that a patient has diabetes, followed by values of 20 unknown features $(\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^{20})$. **Write Python code to train a logistic regression model with stochastic gradient descent (SGD)**. Run SGD for 10,000 iterations, and save the model weights after every 100 iterations. Plot the log-likelihood of the training data given by your model at every 100 iterations. (Log-likelihood is $\log \prod_{i=1}^{n} P(y^i \mid \mathbf{x}^i) = \sum_{i=1}^{n} \log P(y^i \mid \mathbf{x}^i)$ where $(\mathbf{x}^i, y^i)$ is an example.) Provide crystal clear instructions along with the source code on how to execute it. Try a learning rate of 0.1.

From the values of the weights obtained, which five features do you think are the most important? How did you choose these five features?

4. Neural Networks [10 Points]

Derive a stochastic gradient weight update rule for a single unit whose output y is given by

$$y = w_0 + \sum_{i=1}^{n} w_i x_i + w_i x_i^2.$$

We will use the same error functions from lecture notes.

$$E_d = \frac{1}{2} \sum_j (y_j - y_j*)^2.$$

5. Naive Bayes [4 points]

Suppose that we have a problem with two binary inputs, x_1 and x_2, which are truly conditionally independent given the class label. How many parameters does the Naive Bayes classifier have? How many parameters does the classifier with three features have? Generally, how many parameters does Naive Bayes classifier have with n features?