DDA3020: Homework II

October 23, 2022

Homework due: **23:59, November 08, 2022**. The exercise numbers refer to Kevin P. Murphy's book "Machine Learning: A Probabilistic Perspective". The total score of this assignment is 15.

1. Regularizing separate terms in 2d logistic regression **(Exercise 8.7 of Murphy's book)** (1 point)
   (1) Consider the data in Figure 1, where we fit the model

   $$p(y=1 \mid \mathbf{x}, \mathbf{w})=\sigma\left(w_0+w_1 x_1+w_2 x_2\right) .$$

   Suppose we fit the model by maximum likelihood, i.e.,

   $$J(\mathbf{w})=-\ell\left(\mathbf{w}, \mathcal{D}_{\text {train }}\right),$$

   where $\ell\left(\mathbf{w}, \mathcal{D}_{\text {train }}\right)$ is the log likelihood on the training set. Sketch a possible decision boundary corresponding to $\hat{\mathbf{w}}$. (Copy the figure first (a rough sketch is enough), and then superimpose your answer on your copy, since you will need multiple versions of this figure). Is your answer (decision boundary) unique? How many classification errors does your method make on the training set?

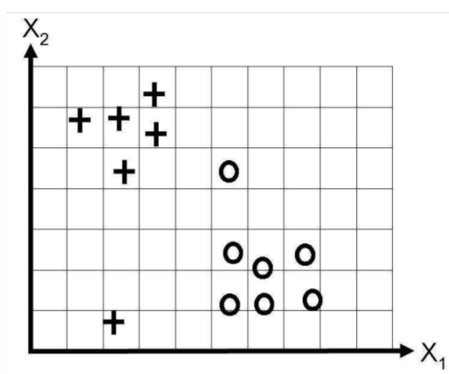

Figure 1: Data for logistic regression question

(2) Now suppose we regularize only the w_0 parameter, i.e., we minimize

$$J_0(\mathbf{w}) = -\ell\left(\mathbf{w}, \mathcal{D}_{\text{train}}\right) + \lambda w_0^2.$$

Suppose λ is a very large number, so we regularize w_0 all the way to 0, but all other parameters are unregularized. Sketch a possible decision boundary. How many classification errors does your method make on the training set? Hint: consider the behavior of simple linear regression, $w_0 + w_1 x_1 + w_2 x_2$ when $x_1 = x_2 = 0$.
(3) Now suppose we heavily regularize only the w_1 parameter, i.e., we minimize

$$J_1(\mathbf{w}) = -\ell\left(\mathbf{w}, \mathcal{D}_{\text{train}}\right) + \lambda w_1^2$$

Sketch a possible decision boundary. How many classification errors does your method make on the training set?
(4) Now suppose we heavily regularize only the w_2 parameter. Sketch a possible decision boundary. How many classification errors does your method make on the training set?

2. Fitting an SVM classifier by hand **(Exercise 14.1 of Murphy's book)** (2 points)
Consider a dataset with 2 points in 1d: $(x_1 = 0, y_1 = -1)$ and $(x_2 = \sqrt{2}, y_2 = 1)$. Consider mapping each point to 3d using the feature vector $\phi(x) = [1, \sqrt{2}x, x^2]^\top$. (This is equivalent to using a second order polynomial kernel.) The max margin classifier has the form

$$\begin{gathered} \min \|\mathbf{w}\|^2 \quad \text{s.t.} \\ y_1\left(\mathbf{w}^T \phi\left(\mathbf{x}_1\right) + w_0\right) \geq 1 \\ y_2\left(\mathbf{w}^T \phi\left(\mathbf{x}_2\right) + w_0\right) \geq 1 \end{gathered}$$

(1) Write down a vector that is parallel to the optimal vector $\mathbf{w}$.
(2) What is the value of the margin that is achieved by this $\mathbf{w}$? Hint: recall that the margin is the distance from each support vector to the decision boundary. Hint 2: think about the geometry of 2 points in space, with a line separating one from the other.
(3) Solve for $\mathbf{w}$, using the fact the margin is equal to $1/\|\mathbf{w}\|$.
(4) Solve for w_0 using your value for $\mathbf{w}$ and the optimization problem above. Hint: the points will be on the decision boundary, so the inequalities will be tight.
(5) Write down the form of the discriminant function $f(x) = w_0 + \mathbf{w}^\top \phi(x)$ as an explicit function of x.

3. Given a binary data set: (2 points)

$$\text{Class -1: } \begin{bmatrix} (1 & 0) \\ (0 & 1) \\ (-1 & 0) \\ (0 & -1) \end{bmatrix} \quad \text{Class } +1: \begin{bmatrix} (2 & 0) \\ (0 & 2) \\ (-2 & 0) \\ (0 & -2) \end{bmatrix}$$

(1) Can you find a svm classifier(without slack variable) for this data set? explain why; (1 point)
(2) Use SVM by expanding the original feature vector $\mathbf{x} = [x_1; x_2]$ to $\phi(\mathbf{x}) = [x_1^2; x_2^2]$, find the svm of this given data set and predict the label of $[1; 2]$. (1 point)

4. Show that the value γ of the margin width for the maximum-margin hyperplane is given by

$$\frac{1}{\gamma^2} = \sum_{n=1}^{N} \alpha_n,$$

where $\{\alpha_n\}$ are given by the following optimization problem (2 points)

$$\begin{aligned}
\max_{\boldsymbol{\alpha}} \quad & \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m \mathbf{x}_n^\top \mathbf{x}_m \\
s.t. \quad & \sum_{n=1}^{N} \alpha_n y_n = 0 \\
& \alpha_n \geq 0 \quad \forall n = 1, 2, ..., N
\end{aligned}$$

Hint: consider the minimum value of the Lagrange function.

Programming

Task description In this problem you are asked to write a program that construct support vector machine models with different kernel functions and slack variable.

Programming preparation Installation of sklearn package: If you want to install package for system default python, use the command

```
pip install sklearn
```

If you want to install package for a specific version of python, say python3.6, use the command

```
python3.6 -m pip install sklearn
```

Installation of jupyter notebook: If you want to install package for system default python, use the command

```
pip install notebook
```

If you want to install package for a specific version of python, say python3.6, use the command

```
python3.6 -m pip install notebook
```

To use the jupyter notebook, navigate to the directory you want and run jupyter notebook. For example, the director of this homework is under the folder

```
~/Downloads/SVM_hw
```

Then the command would be like this

```
cd ~/Downloads/SVM_hw
python3 -m jupyter notebook
```

Then you can click the .ipynb file and begin your editing.

Datasets You are provided with the training and testing datatset (see *train.txt* and *test.txt*), including 120 training data and 30 testing data, respectively. It covers 3 classes, corresponding to setosa, versicolor, virginica. They are derived from the Iris dataset (`https://archive.ics.uci.edu/ml/datasets/iris`), contains 3 classes of 50 instances each, where each class refers to a type of iris plant. Your task is to classify each iris plant as one of the three possible types.

What you should do You should use the SVM function from python sklearn package, which provides different form of SVM function you can use. For multiclass SVM you should use one vs rest strategy. You are recommended to use sklearn.svm.svc() function. You can use numpy for the vector manipulation. For technical report you should state clearly the optimization problem you are solving, how did you derive it, the meaning of different values in the formulation, and some results suitable for presenting in the report (e.g. training error, testing error). The basic form of SVM is given and you don't need to derive this

$$\begin{aligned} &\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \\ &\text{s.t. } 1 - y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right) \le 0, \forall i \end{aligned}$$

1. (2 points) Calculate using standard SVM model (linear separator). Fit your algorithm on the training dataset, then validate your algorithm on testing dataset. Compute the misclassification error of training and testing datasets, the weight vector $\mathbf{w}$, the bias b, and the indices of support vectors(start with 0). Write output to file **SVM_linear.txt**. Note that the sklearn package doesn't provide a function with strict separation so we will simulate this using $C = 1e5$. You should print out the coefficient for each different class separately. The output format should be like this

```
${training_error}
${testing_error}
${w_of_setosa}
```

```
${b_of_setosa}
${support_vector_indices_of_setosa}
${w_of_versicolor}
${b_of_versicolor}
${support_vector_indices_of_versicolor}
${w_of_virginica}
${b_of_virginica}
${support_vector_indices_of_virginica}
```

where each line contains one variable. The training error and testing error count the total error instead of error for each distinct class, the error is $\frac{\text{wrong prediction}}{\text{number of data}}$. If we view the one vs all strategy as combining the multiple different SVM, each one being a separating hyperplane for one class and the rest of the points, then the w, b and support vector indices for that class is the corresponding parameters for the SVM separating this class and the rest of the points. If a variable is of vector form, say $\alpha = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$, then you should write each entry in the same line with comma separated, e.g.

```
1,2,3
```

You should also mention in your report on which classes are linear separable with SVM without slack and how you find it out.

2. (3 points) Calculate using SVM with slack variables. For each $C = 0.1 \times t, t = 1, 2, ..., 10$, fit your algorithm on the training dataset, then validate your algorithm on testing dataset. Compute the misclassification error of training and testing datasets, the weight vector $\mathbf{w}$, the bias b, the indices of support vectors, and the slack variable $\boldsymbol{\xi}$. Write output to file **SVM_slack.txt**. The format is

```
${training_error}
${testing_error}
${w_of_setosa}
${b_of_setosa}
${support_vector_indices_of_setosa}
${slack_variable_of_setosa}
${w_of_versicolor}
${b_of_versicolor}
${support_vector_indices_of_versicolor}
${slack_variable_of_versicolor}
${w_of_virginica}
```

```
${b_of_virginica}
${support_vector_indices_of_virginica}
${slack_variable_of_virginica}
```

3. (3 points) Implement SVM with kernel functions and slack variables. You should experiment with different kernel functions in this task:

   (a) A 2nd-order polynomial kernel, write output to **SVM_poly2.txt**

   (b) A 3rd-order polynomial kernel, write output to **SVM_poly3.txt**

   (c) Radial Basis Function kernel with $\sigma = 1$, write output to **SVM_rbf.txt**

   (d) Sigmoidal kernel with $\sigma = 1$, write output to **SVM_sigmoid.txt**

   During these tasks we set $C = 1$. The output format is

```
${training_error}
${testing_error}
${b_of_setosa}
${support_vector_indices_of_setosa}
${b_of_versicolor}
${support_vector_indices_of_versicolor}
${b_of_virginica}
${support_vector_indices_of_virginica}
```

Note that you should submit A2_StudentID.pdf (report, together with the written answers), A2_StudentID.ipynb (code) and 6 output TXT files (outputs), please zip them into "A2_StudentID.zip".