DDA3020: Homework II

March 26, 2023

Homework due: **23:59, April 09, 2023**. This assignment accounts for 15/100 of the final score

1. Elementary properties of ℓ_2 regularized logistic regression (10 points)
   Consider minimizing

$$J(\mathbf{w}) = -\ell\left(\mathbf{w}, \mathcal{D}_{\text{train}}\right) + \lambda\|\mathbf{w}\|_2^2$$

   where

$$\ell(\mathbf{w}, \mathcal{D}) = \frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}} \log\sigma\left(y_i \mathbf{x}_i^T \mathbf{w}\right)$$

   is the average log-likelihood on data set $\mathcal{D}$, for $y_i \in \{-1, +1\}$.
   Answer the following true/ false questions, and for each question briefly explain why.
   (1) $J(\mathbf{w})$ has multiple locally optimal solutions: T/F?
   (2) Let $\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} J(\mathbf{w})$ be a global optimum. $\hat{\mathbf{w}}$ is sparse (has many zero entries): T/F?
   (3) If the training data is linearly separable, then some weights w_j might become infinite if $\lambda = 0$: T/F?
   (4) $\ell\left(\hat{\mathbf{w}}, \mathcal{D}_{\text{train}}\right)$ always increases as we increase λ : T/F ?
   (5) $\ell\left(\hat{\mathbf{w}}, \mathcal{D}_{\text{test}}\right)$ always increases as we increase λ : T/F ?

2. In class, we saw that if our data is not linearly separable, then we need to modify our support vector machine algorithm by introducing an error margin that must be minimized. In this problem we will consider an method known as l_2 norm soft margin SVM. This new algorithm is given by the following optimization problem (notice that the slack penalties are now squared.) (20 points)

$$\begin{aligned} min_{w,b,\xi} \quad & \tfrac{1}{2}\|w\|^2 + \tfrac{C}{2}\textstyle\sum_{i=1}^m \xi_i^2 \\ s.t. \quad & y^{(i)}(w^T x^{(i)}) + b) \geq 1 - \xi_i, \ i = 1, ..., m \end{aligned}$$

   (1) Notice that we have dropped the $\xi_i \geq 0$ constraint in the l_2 problem. Show that these non-negativity constraints can be removed. That is, show that the optimal value of the objective will be the same whether or not these constraints are present.

(2) What is the Lagrangian of the l_2 soft margin SVM optimization problem $\mathcal{L}(w, b, \xi, \alpha)$? (You can directly give the answer)
(3) Minimize the Lagrangian with respect to w, b, ξ by taking the following gradients: $\frac{\partial \mathcal{L}}{\partial w}$, $\frac{\partial \mathcal{L}}{\partial b}, \frac{\partial \mathcal{L}}{\partial \xi}$ and then setting them to 0. Here, $\xi = [\xi_1, \xi_2, ..., \xi_m]^T$
(4)What is the dual of the l_2 norm soft margin SVM optimization problem with respect to the dual variable α (without the primal variables w, b, and ξ)?

3. Given a binary data set: (20 points)

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

(1) Can you find a svm classifier(without slack variable) for this data set? explain why;
(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 $[(-\frac{1}{\sqrt{2}} \quad \sqrt{2})]$.

4. We will use the dataset below to learn a decision tree which predicts if people pass machine learning (Yes or No), based on their previous GPA (High, Medium, or Low) and whether or not they studied. For this problem, you can write your answers using $\log_2$, but it may be helpful to note that $\log_2 3 \approx 1.6$. (15 points)

GPA	Studied	Passed
L	F	F
L	T	T
M	F	F
M	T	T
H	F	T
H	T	T

(1) What is the entropy H(Passed)?
(2) What is the entropy H(Passed | GPA)?
(3) What is the entropy H(Passed | Studied)?
(4) Draw the full decision tree that would be learned for this dataset. You do not need to show any calculations.

Programming (50 points)

Warm-up of SVM (5 points) There are some basic illustration and warm-up questions in the provided SVM.ipynb files on BB. Please read

them and finish these questions before you start the formal examples about wine analysis.

You do not need to write reports about this part. Just finish it on the ipynb files.

Wine Analysis with SVM (45 points) In this part you are asked to write a program that construct support vector machine models with different kernel functions and slack variable.

Datasets You need to import the 'Wine Analysis' datasets from sklearn.

You can use 'sklearn.datasets.load_wine()'.

The datasets includes 178 instances and 13 attributes. It covers 3 classes with class distribution: class_0 (59 datapoints), class_1 (71 datapoints), class_2 (48 datapoints). This is a copy of UCI ML Wine recognition datasets. `https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data`. Your task is to classify each sample of wine 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.

Report For technical report of each subproblem as below, you should state clearly:

1. the optimization problem you are solving, and how did you derive it; Hint: you can refer to the slides and make a sketch proof breifly.

2. the meaning of different values in the formulation, and some results suitable for presenting in the report (e.g. training error, testing error, the indices of support vectors, etc).

3. analysis how the parameter settings influence the final results (if any).

The basic form of SVM is given and you don't need to derive this

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

(a) (15 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).

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.

Report the experiments results in your pdf files like this, you can use table/snapshot or any feasible method to demonstrate these data.

Attention: please demonstrate this part in your report. If you only show it in ipynb files, you will not get the marks.

```
${training_error}
${testing_error}
${w_of_wine_type_0}
${b_of_wine_type_0}
${support_vector_indices_of_wine_type_0}
${w_of_wine_type_1}
${b_of_wine_type_1}
${support_vector_indices_of_wine_type_1}
${w_of_wine_type_2}
${b_of_wine_type_2}
${support_vector_indices_of_wine_type_2}
```

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.

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

(b) (15 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 $\boldsymbol{\xi}$. The results you need to report are

```
${training_error}
${testing_error}
${w_of_wine_type_0}
${b_of_wine_type_0}
${support_vector_indices_of_wine_type_0}
${slack_variable_of_wine_type_0}
${w_of_wine_type_1}
```

```
${b_of_wine_type_1}
${support_vector_indices_of_wine_type_1}
${slack_variable_of_wine_type_1}
${w_of_wine_type_2}
${b_of_wine_type_2}
${support_vector_indices_of_wine_type_2}
${slack_variable_of_wine_type_2}
```

(c) (15 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;

(b) A 3rd-order polynomial kernel;

(c) Radial Basis Function kernel with $\sigma = 1$;

(d) Sigmoidal kernel with $\sigma = 1$;

During these tasks we set $C = 1$. The results' format of each kernel functions is

```
${training_error}
${testing_error}
${b_of_wine_type_0}
${support_vector_indices_of_wine_type_0}
${b_of_wine_type_1}
${support_vector_indices_of_wine_type_1}
${b_of_wine_type_2}
${support_vector_indices_of_wine_type_2}
```

Note that you should submit A2_StudentID.pdf (report, together with the written answers), A2_StudentID.ipynb (code).