Foundations of Machine Learning: Assignment 2

1 Perceptron

Suppose we have N linearly separable points $\{\boldsymbol{x}_i, i = 1, \ldots, N\}$ in $\mathbb{R}^d$, with class labels $y_i \in \{-1, 1\}, i = 1, \ldots, N$. Show that the perceptron learning algorithm converges to a separating hyperplane in a finite number of steps:
(a) Denoting a hyperplane by $\boldsymbol{a}^T\boldsymbol{x} + b = 0$, or in more compact notation $\boldsymbol{w}^T\bar{\boldsymbol{x}} = 0$, where $\bar{\boldsymbol{x}}^T = (\boldsymbol{x}^T, 1)$ and $\boldsymbol{w}^T = (\boldsymbol{a}^T, b)$. Let $\boldsymbol{z}_i = \frac{\bar{\boldsymbol{x}}}{||\bar{\boldsymbol{x}}||}$. Show that separability implies the existence of a $\hat{\boldsymbol{w}}$ such that

$$y_i\hat{\boldsymbol{w}}^T\boldsymbol{z}_i \geq 1, \forall i \in \{1, \ldots, N\}$$

(b) Given a current $\boldsymbol{w}_t$, the perceptron algorithm identifies a point $\boldsymbol{z}_i$ that is misclassified, and produces the update $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + y_i\boldsymbol{z}_i$. Show that $||\boldsymbol{w}_{t+1} - \hat{\boldsymbol{w}}||^2 \leq ||\boldsymbol{w}_t - \hat{\boldsymbol{w}}||^2 - 1$, and hence the algorithm converges to a separating hyperplane in $\lceil ||\boldsymbol{w}_0 - \hat{\boldsymbol{w}}||^2 \rceil$ steps. (Where $\hat{\boldsymbol{w}}$ is the parameter of a separating hyperplane)

2 Regression and Linear Classification

2.1 Linear Regression: Closed-form Solution and Geometrical Interpretation

Suppose we have N data points $\{\boldsymbol{x}_i \in \mathbb{R}^d, i = 1, \ldots, N\}$, with corresponding targets $y_i \in \mathbb{R}, i = 1, \ldots, N$. The linear regression problem aims to find the optimal $\boldsymbol{w}, b$ such that

$$\boldsymbol{w}, b = \arg\min_{\boldsymbol{w},b} F(\boldsymbol{w}, b) = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{w}^T\boldsymbol{x}_i + b - y_i)^2.$$

Rewrite the objective function as

$$F(\boldsymbol{W}) = \frac{1}{N}||\boldsymbol{X}^T\boldsymbol{W} - \boldsymbol{Y}||^2,$$

where $\boldsymbol{X}^T = \begin{bmatrix} \boldsymbol{x}_1^T & 1 \\ \vdots & \\ \boldsymbol{x}_N^T & 1 \end{bmatrix}$, $\boldsymbol{W}^T = [\boldsymbol{w}^T, b]$ and $\boldsymbol{Y}^T = [y_1, y_2, \ldots, y_N]$
(a) show that the optimization problem has closed-form solution. (suppose $\boldsymbol{X}\boldsymbol{X}^T$ is invertible.)

$$\boldsymbol{W} = (\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}\boldsymbol{Y}$$

(b) Show that the matrix $\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}$ takes any vector $\mathbf{v}$ and projects it onto the space spanned by the columns of $\boldsymbol{X}^T$. Use this result to show that the least-squares solution in (1) corresponds to an orthogonal projection of the vector $\boldsymbol{Y}$ onto the subspace $\mathcal{S}$ as shown in Figure 1 (i.e. show that $\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}\boldsymbol{Y} - \boldsymbol{Y}$ is orthogonal to the space spanned by the columns of $\boldsymbol{X}^T$).

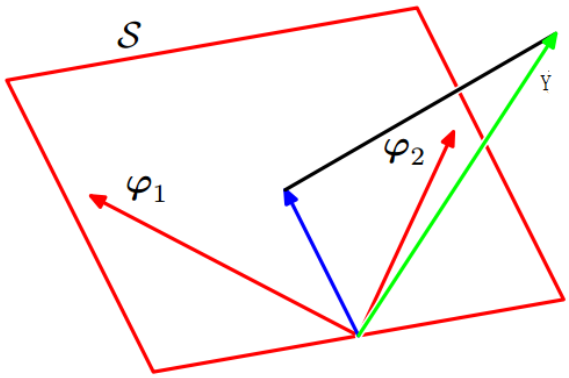

Figure 1

2.2 Logistic Regression: Overfitting Problem

Show that for a linearly separable data set, the maximum likelihood solution for the logistic regression model is obtained by finding a vector $\boldsymbol{w}$ whose decision boundary $\boldsymbol{w}^T\boldsymbol{x} = 0$ separates the classes and then taking the magnitude of $\boldsymbol{w}$ to infinity.

2.3 Fisher's Linear Discriminant: Relation to Least Squares

Using the definitions of the between-class and within-class covariance matrices

$$S_B = (\boldsymbol{m}_2 - \boldsymbol{m}_1)(\boldsymbol{m}_2 - \boldsymbol{m}_1)^T$$
$$S_W = \sum_{i\in\mathcal{C}_1}(\boldsymbol{x}_i - \boldsymbol{m}_1)(\boldsymbol{x}_i - \boldsymbol{m}_1)^T + \sum_{i\in\mathcal{C}_2}(\boldsymbol{x}_i - \boldsymbol{m}_2)(\boldsymbol{x}_i - \boldsymbol{m}_2)^T$$

respectively, together with

$$w_0 = -\boldsymbol{w}^T\boldsymbol{m}$$
$$\boldsymbol{m} = \frac{1}{M}(M_1\boldsymbol{m}_1 + M_2\boldsymbol{m}_2)$$

and the choice of target values described in Lecture, i.e.

$$y_i = \begin{cases} \dfrac{M}{M_1} & i \in \mathcal{C}_1 \\ \dfrac{M}{M_2} & i \in \mathcal{C}_2 \end{cases}$$

Show that the expression

$$\sum_{i=1}^{M}(\boldsymbol{w}^T\boldsymbol{x}_i + w_0 - y_i)\boldsymbol{x}_i = 0$$

that minimizes the sum-of-squares error function can be written in the form

$$\left(S_W + \frac{M_1M_2}{M}S_B\right)\boldsymbol{w} = M(\boldsymbol{m}_1 - \boldsymbol{m}_2)$$

3 SVM

(Part of exercise 4.1 in *Foundations of Machine Learning*) One can give two types of arguments in favor of the SVM algorithm: one based on the sparsity of the support vectors, another based on the notion of margin.

Suppose that instead of maximizing the margin, we choose instead to maximize sparsity by minimizing the L_p norm of the vector $\boldsymbol{\alpha}$ that defines the weight vector $\boldsymbol{w}$, for some $p \geq 1$. First, consider the case $p = 2$. This gives the following optimization problem:

$$\begin{aligned} \min_{\boldsymbol{\alpha},b} \quad & \frac{1}{2}\sum_{i=1}^{m}\alpha_i^2 + C\sum_{i=1}^{m}\xi_i \\ \text{subject to} \quad & y_i\left(\sum_{j=1}^{m}\alpha_j y_j \boldsymbol{x}_i^T \boldsymbol{x}_j + b\right) \geq 1 - \xi_i, i \in [1,m]. \\ & \xi_i, \alpha_i \geq 0, i \in [1,m]. \end{aligned} \tag{1}$$

(a) Derive the dual optimization of problem of optimization problem (1).
(b) Setting $p = 1$ will induce a more sparse $\boldsymbol{\alpha}$. Derive the dual optimization in this case.

4 Program Practice

In this problem, you are supposed to solve a binary classification problem with logistic regression and SVM, then comparing the results. The dataset to be used is *UCI: Pima Indians Diabetes Data Set*, which is included in the attachment. From http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes you can acquire the dataset and its illustration.

Logistic Regression

To do:

(a) Apply gradient ascent method to optimize your vanilla logistic regression model.
(b) Introduce regularization to logistic regression. Choose the optimal regularization parameter by cross-validation.
(c) (Optional, bonus) Apply the IRLS method to vanilla and regularized logistic regression model.(IRLS method is introduced in attachment)

To Report:

You are required to include the following results in your report:
(a) Explain how do you split the dataset into training set, validation set and test set.
(b) Report your choice of regularization parameters and how do you make your decision.
(c) Show the typical curves, for example the curves of loss and accuracy. Report your final results which are better shown in mean±std format.
(d) (Optional, bonus) Show your results of IRLS method and compare the convergence speed with gradient ascent method.

SVM

To do:

(a) We recommend you to use the library *libsvm*: http://www.csie.ntu.edu.tw/~cjlin/libsvm/. Learn to use it by yourself and then apply to this problem.
(b) There are a lot of hyper-parameters you need to choose such as kernel function and regularization parameters. You need to choose them by theoretical consideration or experiments (cross-validation).

To Report:

You are required to include the following results in your report:
(a) Report how do you choose hyper-parameters and your final choice.
(b) Show your results.

Notes: We have no restriction on the program language, you can choose your favorite program language. But we think the **MATLAB** or **Python** may be easier and we recommend you to use them.