Exercise no 7

FirstName LastName StudentNumber

Statistical Machine Learning

May 6, 2018

1 Mathematical Statistics

Exercise 1.1 Take a look at the objective function of linear support vector machines $D(\beta_0, \boldsymbol{\beta})$ from "The elements of Statistical Learning" (ESL) page 131.

Show that minimizing

$$D(\beta_0, \boldsymbol{\beta}) = -\sum_{i \in \mathcal{M}} y_i(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta})$$

is equivalent to minimizing

$$S(\beta_0, \boldsymbol{\beta}) = \sum_{i=1}^{n} \{1 - y_i(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta})\}_+.$$

Hint: Start with finding $(\beta_0, \boldsymbol{\beta})$ so that $y_i(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}) > 0, \forall i = 1, \ldots, n$. Since n is finite it is equivalent to $y_i(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}) > \delta$ for some $\delta > 0$. (A trivial choice of $0 < \delta < \min_{i=1,\ldots,n}\{y_i(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta})\}$, which means the vector $(\beta_0, \boldsymbol{\beta})$ is re-scalable.) This means observation i is misclassified if $y_i(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}) < \delta$ and now argue it is the same as minimizing

$$S(\beta_0, \boldsymbol{\beta}) = \sum_{i=1}^{n} \{\delta - y_i(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta})\}_+$$

To avoid re-scalability problem, choose for instance $\delta = 1$. Still the minimizer $(\hat{\beta}_0, \hat{\boldsymbol{\beta}})$ is non-unique if data are separable, because any $S(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) = 0$ is a solution (if you do not understand why you encounter uniqueness problem, look at Figure 4.14 page 129 of ESL giving two blue lines both with $S(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) = 0$; this uniqueness problem is the same as the convergence problem of the logistic regression while data are separable).

Connection to Ridge Regression

One way to find a unique solution is by adding L_2 penalty for a given $\lambda > 0$ or so called

maximizing the margin

$$S(\beta_0, \boldsymbol{\beta}) = \sum_{i=1}^{n} \{1 - y_i(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta})\}_+ + \lambda \sum_{j=1}^{p} \beta_j^2$$

A special choice of λ while data are not separable gives the maximum margin support vector classifier. This suggests the ridge estimator is a good initial value for the SVM! You can show that the hinge loss above resembles logistic regression objective function, so SVM is like L2 penalized logistic regression.

Solution 1.1 Put your proof here.

2 Application

Exercise 2.1 Create a decision tree of depth 5 for the spam data shared on the course website. You may use the `classtree.ipynb` code shared on the course website and play with it.

Solution 2.1 Put the tree here.