Learning From Data
Caltech
http://work.caltech.edu/telecourse.html
2013

Online Homework # 8

Collaboration in the sense of discussions is allowed, but you should NOT discuss your selected answers with anyone. Books and notes can be consulted. All questions will have multiple choice answers ([**a**], [**b**], [**c**], **...***). You should enter your solutions online by logging into your account at the course web site.*

Note about the homeworks

- The goal of the homeworks is to facilitate a deeper understanding of the course material. The questions are not designed to be puzzles with catchy answers. They are meant to make you roll your sleeves, face uncertainties, and approach the problem from different angles.
- The problems range from easy to hard, and from theoretical to practical. Some problems require running a full experiment to arrive at the answer.
- The answer may not be obvious or numerically close to one of the choices. The intent is to prompt discussion and exchange of ideas.
- Speaking of discussion, you are encouraged to take part in the forum

  http://book.caltech.edu/bookforum

  where there are many threads about each homework. We hope that you will contribute to the discussion as well.
- Please follow the forum guidelines for posting answers (see the "BEFORE posting answers" announcement at the top there).

Primal versus Dual Problem

1. Recall that N is the size of the data set and d is the dimensionality of the input space. The original formulation of the hard-margin SVM problem (minimize $\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$ subject to the inequality constraints), without going through the Lagrangian dual problem, is

[**a**] a quadratic programming problem with N variables

[**b**] a quadratic programming problem with $N+1$ variables

[**c**] a quadratic programming problem with d variables

[**d**] a quadratic programming problem with $d+1$ variables

[**e**] not a quadratic programming problem

Support Vector Machines With Soft Margins

In this homework set, we are going to experiment with a real-world dataset. Download the processed US Postal Service Zip Code dataset with extracted features of symmetry and intensity for training and testing:

http://www.amlbook.com/data/zip/features.train

http://www.amlbook.com/data/zip/features.test

(the format of each row is: **digit symmetry intensity**). We will train two types of binary classifiers; one-versus-one (one digit is class +1 and another digit is class −1, with the rest of the digits disregarded), and one-versus-all (one digit is class +1 and the rest of the digits are class −1).

The data set has thousands of points, and some quadratic programming packages cannot handle this size. We recommend that you use the packages in libsvm:

http://www.csie.ntu.edu.tw/∼cjlin/libsvm/

Implement SVM with soft margin on the above zip-code data set by solving

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\alpha_n\alpha_m y_n y_m K(\mathbf{x}_n,\mathbf{x}_m) - \sum_{n=1}^{N}\alpha_n \\ \text{s.t.} \quad & \sum_{n=1}^{N} y_n\alpha_n = 0 \\ & 0 \le \alpha_n \le C \quad n = 1,\cdots,N \end{aligned}$$

When evaluating E_{in} and E_{out} of the resulting classifier, use binary classification error.

Practical remarks:

(i) For the purpose of this homework, do not scale the data when you use libsvm or other packages, lest you should change the effective kernel and get different results.

(ii) In some packages, you need to specify double precision.

(iii) In 10-fold cross validation, if the data size is not a multiple of 10, the sizes of the 10 subsets may be off by 1 data point.

(iv) Some packages have software parameters whose values affect the outcome. ML practitioners have to deal with this kind of added uncertainty.

Polynomial Kernels

Consider the polynomial kernel $K(\mathbf{x}_n, \mathbf{x}_m) = (1 + \mathbf{x}_n^{\mathrm{T}}\mathbf{x}_m)^Q$, where Q is the degree of the polynomial.

2. With $C = 0.01$ and $Q = 2$, which of the following classifiers has the **highest** E_{in}?

[**a**] 0 versus all

[**b**] 2 versus all

[**c**] 4 versus all

[**d**] 6 versus all

[**e**] 8 versus all

3. With $C = 0.01$ and $Q = 2$, which of the following classifiers has the **lowest** E_{in}?

[**a**] 1 versus all

[**b**] 3 versus all

[**c**] 5 versus all

[**d**] 7 versus all

[**e**] 9 versus all

4. Comparing the two classifiers from Problems 2 and 3, which of the following values is the closest to the difference between the number of support vectors of these two classifiers?

[**a**] 600

[**b**] 1200

[**c**] 1800

[**d**] 2400

[**e**] 3000

5. Consider the 1 versus 5 classifier with $Q = 2$ and $C \in \{0.001, 0.01, 0.1, 1\}$. Which of the following statements is correct?

[**a**] The number of support vectors goes down when C goes up

[**b**] The number of support vectors goes up when C goes up

[**c**] E_{out} goes down when C goes up

[**d**] Maximum C achieves the lowest E_{in}

[**e**] None of the above

6. In the 1 versus 5 classifier, comparing $Q = 2$ with $Q = 5$, which of the following statements is correct?

[**a**] When $C = 0.0001$, E_{in} is higher at $Q = 5$

[**b**] When $C = 0.001$, the number of support vectors is lower at $Q = 5$

[**c**] When $C = 0.01$, E_{in} is higher at $Q = 5$

[**d**] When $C = 1$, E_{out} is lower at $Q = 5$

[**e**] None of the above

Cross Validation

In the next two problems, we will experiment with 10-fold cross validation for the polynomial kernel. Because E_{cv} is a random variable that depends on the random partition of the data, we will try 100 runs with different partitions, and base our answer on the number of runs that lead to a particular choice.

7. Consider the 1 versus 5 classifier with $Q = 2$. We use E_{cv} to select $C \in \{0.0001, 0.001, 0.01, 0.1, 1\}$. If there is a tie in E_{cv}, select the smaller C. Within the 100 random runs, which of the following statements is correct?

[**a**] $C = 0.0001$ is selected most often

[**b**] $C = 0.001$ is selected most often

[**c**] $C = 0.01$ is selected most often

[**d**] $C = 0.1$ is selected most often

[**e**] $C = 1$ is selected most often

8. Again, consider the 1 versus 5 classifier with $Q = 2$. For the winning selection in the previous problem, the average value of E_{cv} over the 100 runs is closest to

(a) 0.001

(b) 0.003

(c) 0.005

(d) 0.007

(e) 0.009

RBF Kernel

Consider the radial basis function (RBF) kernel $K(\mathbf{x}_n, \mathbf{x}_m) = \exp(-||\mathbf{x}_n - \mathbf{x}_m||^2)$. Focus on the 1 versus 5 classifier.

9. Which of the following values of C results in the lowest E_{in}?

[**a**] $C = 0.01$

[**b**] $C = 1$

[**c**] $C = 100$

[**d**] $C = 10^4$

[**e**] $C = 10^6$

10. Which of the following values of C results in the lowest E_{out}?

[**a**] $C = 0.01$

[**b**] $C = 1$

[**c**] $C = 100$

[**d**] $C = 10^4$

[**e**] $C = 10^6$