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

Online Homework # 7

*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).

Validation

In the following problems, use the data provided in the files `in.dta` and `out.dta` for Homework #6. We are going to apply linear regression with a nonlinear transformation for classification (without regularization). The nonlinear transformation is given by ϕ_0 through ϕ_7 which transform (x_1, x_2) into

$$1 \qquad x_1 \qquad x_2 \qquad x_1^2 \qquad x_2^2 \qquad x_1x_2 \qquad |x_1 - x_2| \qquad |x_1 + x_2|$$

To illustrate how taking out points for validation affects the performance, we will consider the hypotheses trained on $\mathcal{D}_{\text{train}}$ (without restoring the full $\mathcal{D}$ for training after validation is done).

1. Split `in.dta` into training (first 25 examples) and validation (last 10 examples). Train on the 25 examples only, using the validation set of 10 examples to select between five models that apply linear regression to ϕ_0 through ϕ_k, with $k = 3, 4, 5, 6, 7$. For which model is the classification error on the validation set smallest?

[**a**] $k = 3$

[**b**] $k = 4$

[**c**] $k = 5$

[**d**] $k = 6$

[**e**] $k = 7$

2. Evaluate the out-of-sample classification error using `out.dta` on the 5 models to see how well the validation set predicted the best of the 5 models. For which model is the out-of-sample classification error smallest?

[**a**] $k = 3$

[**b**] $k = 4$

[**c**] $k = 5$

[**d**] $k = 6$

[**e**] $k = 7$

3. Reverse the role of training and validation sets; now training with the (last 10 examples and validating with the first 25 examples. For which model is the classification error on the validation set smallest?

[**a**] $k = 3$

[**b**] $k = 4$

[**c**] $k = 5$

[**d**] $k = 6$

[**e**] $k = 7$

4. Once again evaluate the out-of-sample classification error using `out.dta` on the 5 models to see how well the validation set predicted the best of the 5 models. For which model is the out-of-sample classification error smallest?

[**a**] $k = 3$

[**b**] $k = 4$

[**c**] $k = 5$

[**d**] $k = 6$

[**e**] $k = 7$

5. What values are closest to the out-of-sample classification error obtained for the model chosen in each of the above two experiments, respectively?

[**a**] 0.0, 0.1

[**b**] 0.1, 0.2

[**c**] 0.1, 0.3

[**d**] 0.2, 0.2

[**e**] 0.2, 0.3

Estimators

6. Let e_1 and e_2 be independent random variables, distributed uniformly over the interval $[0, 1]$. Let $\mathsf{e} = \min(\mathsf{e}_1, \mathsf{e}_2)$. The expected values of $\mathsf{e}_1, \mathsf{e}_2, \mathsf{e}$ are closest to

[**a**] 0.5, 0.5, 0

[**b**] 0.5, 0.5, 0.1

[**c**] 0.5, 0.5, 0.25

[**d**] 0.5, 0.5, 0.4

[**e**] 0.5, 0.5, 0.5

Cross Validation

7. You are given the data points: $(-1,0), (\rho,1), (1,0)$, $\rho \geq 0$, and a choice between two models: constant $[\,h_0(x) = b\,]$ and linear $[\,h_1(x) = ax + b\,]$. For which value of ρ would the two models be tied using leave-one-out cross-validation with the squared error measure?

[**a**] $\sqrt{\sqrt{3}+4}$

[**b**] $\sqrt{\sqrt{3}-1}$

[**c**] $\sqrt{9+4\sqrt{6}}$

[**d**] $\sqrt{9-\sqrt{6}}$

[**e**] None of the above

PLA vs. SVM

In the following problems, we compare PLA to SVM with hard margin on linearly separable data sets. For each run, you will create your own target function f and data set $\mathcal{D}$. Take $d = 2$ and choose a random line in the plane as your target function f (do this by taking two random, uniformly distributed points on $[-1,1] \times [-1,1]$ and taking the line passing through them), where one side of the line maps to $+1$ and the other maps to -1. Choose the inputs $\mathbf{x}_n$ of the data set as random points in $\mathcal{X} = [-1,1] \times [-1,1]$, and evaluate the target function on each $\mathbf{x}_n$ to get the corresponding output y_n. If all data points are on one side of the line, discard the run and start a new run.

Start PLA with the all-zero vector and pick the misclassified point for each PLA iteration at random. Run PLA to find the final hypothesis g_{PLA} and measure the difference between f and g_{PLA} as $\mathbb{P}[f(\mathbf{x}) \neq g_{\text{PLA}}(\mathbf{x})]$ (you can either calculate this exactly, or approximate it by generating a sufficiently large separate set of points to evaluate it). Now, run SVM on the same data to find the final hypothesis g_{SVM} by solving

$$\begin{array}{ll} \displaystyle\min_{\mathbf{w},b} & \dfrac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \\ \text{s.t.} & y_n\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b\right) \geq 1 \end{array}$$

using quadratic programming on the primal or the dual problem. Measure the difference between f and g_{SVM} as $\mathbb{P}[f(\mathbf{x}) \neq g_{\text{SVM}}(\mathbf{x})]$, and count the number of support vectors you get in each run.

8. For $N = 10$, repeat the above experiment for 1000 runs. How often is g_{SVM} better than g_{PLA} in approximating f? The percentage of time is closest to:

[a] 20%

[b] 40%

[c] 60%

[d] 80%

[e] 100%

9. For $N = 100$, repeat the above experiment for 1000 runs. How often is g_{SVM} better than g_{PLA} in approximating f? The percentage of time is closest to:

[a] 10%

[b] 30%

[c] 50%

[d] 70%

[e] 90%

10. For the case $N = 100$, which of the following is the closest to the average number of support vectors of g_{SVM} (averaged over the 1000 runs)?

[a] 2

[b] 3

[c] 5

[d] 10

[e] 20