Homework 7

Sara Beery
CS 156A - Learning Systems

November 15, 2018

1 Validation

In the following problems we use the data provided in the files

http://work.caltech.edu/data/in.dta
http://work.caltech.edu/data/out.dta

as a training and test set respectively. Each line of the files corresponds to a two-dimensional input $x = (x_1, x_2)$, so that $X = \mathbb{R}^2$, followed by the corresponding label from $Y = \{1, 1\}$. We are going to apply Linear Regression with a non-linear transformation for classification. The nonlinear transformation is given by

$$\Phi(x_1, x_2) = (1, x_1, x_2, x_1^2, x_2^2, x_1x_2, |x_1x_2|, |x_1 + x_2|).$$

Where we denote $\Phi_3(x_1, x_2) = (1, x_1, x_2, x_1^2)$, $\Phi_4(x_1, x_2) = (1, x_1, x_2, x_1^2, x_2^2)$, etc.

After splitting the training set into training (first 25 examples) and validation (last 10 examples) and applying Linear Regression for $k \in \{3, 4, 5, 6, 7\}$, we find that the validation classification error is smallest for $k = 6$, where it equals 0.

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

The out of sample classification error is smallest for $k = 7$, where it equals 0.072

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

Swapping the training and validation sets, we find $k = 6$ still gives us the smallest validation error, with a value of 0.08.

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

In this case we find that the smallest test error also occurs at $k = 6$, with a value of 0.192

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

The out of sample errors for the models chosen in problem 1 and 3 are $(0.084, 0.192)$ which is closest to:

5. **[b]** $(0.1, 0.2)$

2 Validation Bias

Let e_1 and e_2 be independent random variables, distributed uniformly over the interval $[0, 1]$. Let $e = min(e_1, e_2)$. The expected values of e_1 and e_2 are 0.5. In order to find the expected value of e, first note that $\mathbb{E}(e) = \int_0^1 x f_e(x) dx$ where $f_e(x)$ is the density function.

In order to find the density function for e, we recall that the density function is the derivative with respect to x of the cumulative distribution function $F_e(x) = \mathbb{P}[e \leq x] = \mathbb{P}[min(e_1, e_2) \leq x]$. Note that

$$\begin{aligned}
\mathbb{P}[min(e_1, e_2) \leq x] &= 1 - \mathbb{P}[min(e_1, e_2) \geq x] \\
&= 1 - \mathbb{P}[e_1 \geq x, e_2 \geq x] \\
&= 1 - \mathbb{P}[[e_1 \geq x]\mathbb{P}[[e_2 \geq x] \\
&= 1 - \mathbb{P}[[e_1 \geq x]^2 \\
&= 1 - (1 - \mathbb{P}[[e_1 \leq x])^2 \\
&= 1 - (1 - x)^2,
\end{aligned}$$

and therefore

$$f_e(x) = 2(1 - x)$$

and at last we can calculate

$$\mathbb{E}(e) = 2 \int_0^1 x(1 - x) dx = \frac{1}{3} \approx 0.3333.$$

This is closest to

6. **[d]** $0.5, 0.5, 0.4$

3 Cross Validation

Given the data points $(x, y) : (1, 0), (\rho, 1), (1, 0), \rho 0$, and a choice between two models: constant $\{h_0(x) = b\}$ and linear $\{h_1(x) = ax + b\}$, we do leave-one-out cross validation with the squared error measure. This means that for each held-out point, we will calculate the best model of each type of the remaining points. We will then calculate the error on the held out point. The error for the model will be the sum of the associated squared errors on the held-out points.

Holding out $(x_1, y_1) = (-1, 0)$, we find the best fit for the points $(1, 0)$ and $(\rho, 1)$. For the constant model, we get

$$g_0^{(1)}(x) = 0.5.$$

For the linear model, we get

$$g_1^{(1)}(x) = \frac{x+1}{\rho+1}.$$

Holding out $(x_2, y_2) = (1, 0)$, we find the best fit for the points $(-1, 0)$ and $(\rho, 1)$. For the constant model, we get

$$g_0^{(2)}(x) = 0.5.$$

For the linear model, we get

$$g_1^{(2)}(x) = \frac{x-1}{\rho-1}.$$

Holding out $(x_3, y_3) = (\rho, 1)$, we find the best fit for the points $(-1, 0)$ and $(1, 0)$. For the constant model, we get

$$g_0^{(3)}(x) = 0.$$

For the linear model, we get

$$g_1^{(3)}(x) = 0.$$

Summing the squared error for the constant model, we get

$$e_0 = \sum_{n=1}^{3} (g_0^{(n)}(x_n) - y_n)^2 = \frac{1}{4} + \frac{1}{4} + 1 = \frac{3}{2}$$

Summing the squared error for the linear model, we get

$$e_1 = \sum_{n=1}^{3} (g_1^{(n)}(x_n) - y_n)^2 = \frac{4}{(\rho-1)^2} + \frac{4}{(\rho+1)^2} + 1$$

Setting $e_0 = e_1$ and solving for ρ, we find that the real solutions are $\rho = \sqrt{9+4\sqrt{6}}, -\sqrt{9+4\sqrt{6}}$.

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

4 PLA vs. SVM

When comparing PLA and SVM, we find that when using $N = 10$ we get better results from SVM about 37% of the time, which is closest to:

8. [b] 40%

When using $N = 100$ we get better results from SVM about 27% of the time, which is closest to:

9. [**b**] 25%

We find that with $N = 100$ we get on average 3.001 support vectors for each run of SVM.

10. [**b**] 3