DSC 255: Machine learning

Week 4 — Solutions

Worksheet

1. *Regression with one predictor variable*

   (a) We will predict the mean of the y-values: $\widehat{y} = (1+3+4+6)/4 = 3.5$. The MSE of this prediction is exactly the variance of the y-values, namely:

   $$\text{MSE} = \frac{(1-3.5)^2 + (3-3.5)^2 + (4-3.5)^2 + (6-3.5)^2}{4} = 3.25.$$

   (b) If we simply predict x, the MSE is

   $$\frac{1}{4}\sum_{i=1}^{4}(y^{(i)} - x^{(i)})^2 = \frac{1}{4}\left((1-1)^2 + (1-3)^2 + (4-4)^2 + (4-6)^2\right) = 2.$$

   (c) We saw in class that the MSE is minimized by choosing

   $$a = \frac{\sum_i (y^{(i)} - \overline{y})(x^{(i)} - \overline{x})}{\sum_i (x^{(i)} - \overline{x})^2}$$
   $$b = \overline{y} - a\overline{x}$$

   where $\overline{x}$ and $\overline{y}$ are the mean values of x and y, respectively. This works out to $a = 1, b = 1$; and thus the prediction on x is simply $x + 1$. The MSE of this predictor is:

   $$\frac{1}{4}\left(1^2 + 1^2 + 1^2 + 1^2\right) = 1.$$

2. *Optimality of the mean.*

   (a) $dL/ds = -2(x_1 + \cdots + x_n)/n + 2s$.

   (b) Setting $dL/ds = 0$, we get $s = (x_1 + \cdots + x_n)/n$.

3. We would write the loss induced by a linear predictor $w \cdot x + b$ as

   $$L(w, b) = \sum_{i=1}^{n} |y^{(i)} - (w \cdot x^{(i)} + b)|.$$

4. *Inherent uncertainty.* This is somewhat subjective, but (b), (d) seem pretty clear-cut cases where perfect predictions are not possible.

5. *Logistic regression.*

(a) The decision boundary is the hyperplane given by $c = 1/2$.

(b) $c = 3/4$ yields a hyperplane that is parallel to the decision boundary.

(c) $c = 1/4$ yields a hyperplane parallel to the decision boundary. It is on the opposite side to (b) and the same distance from the decision boundary.

6. *Discovering relevant features in regression.*

   (a) A sensible strategy is to do linear regression using the Lasso, and to choose a regularization constant λ that yields roughly 10 non-zero coefficients.

   (b) First value of λ which gave nonzero coefficients only for 10 features is 0.4. This yielded the following features (numbering starting at 1): $2, 3, 5, 7, 11, 13, 17, 19, 23, 27$.

7. *Binary logistic regression.* See the accompanying notebook `heart-soln.ipynb`. The results obtained depend on the random partition of the data into training and test sets. In one particular run, we got the following results.

   (a) Coefficients:

   $$[0.014, -0.928, 0.588, -0.010, -0.004, -0.269, 0.358, 0.027, -0.869, -0.685, 0.285, -0.736, -0.567]$$

   The three features that were most influential in the model – the features with the highest absolute values – were 1 (`sex`), 8 (`exang`), and 11 (`ca`).

   (b) The test error was 18.45%.

   (c) The 5-fold cross validation error was 16.00%, which is fairly close to the test error.