DSC 255: Machine learning

Homework 4

Regression

1. *Example of regression with one predictor variable.* Consider the following simple data set of four points (x, y):

$$(1, 1), (1, 3), (4, 4), (4, 6).$$

   (a) Suppose you had to predict y without knowledge of x. What value would you predict? What would be its mean squared error (MSE) on these four points?

   (b) Now let's say you want to predict y based on x. What is the MSE of the linear function $y = x$ on these four points?

   (c) Find the line $y = ax + b$ that minimizes the MSE on these points. What is its MSE?

2. *Optimality of the mean.* One fact that we used implicitly in the lecture is the following:

   > If we want to summarize a bunch of numbers $x_1, \ldots, x_n$ by a single number s, the best choice for s, the one that minimizes the average squared error, is the **mean** of the x_i's.

   Let's see why this is true. We begin by defining a suitable loss function. Any value $s \in \mathbb{R}$ induces a mean squared loss (MSE) given by:

$$L(s) = \frac{1}{n} \sum_{i=1}^{n} (x_i - s)^2.$$

   We want to find the s that minimizes this function.

   (a) Compute the derivative of $L(s)$.

   (b) What value of s is obtained by setting the derivative dL/ds to zero?

3. We have a data set $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$, where $x^{(i)} \in \mathbb{R}^d$ and $y^{(i)} \in \mathbb{R}$. Suppose that we want to express y as a linear function of x, but the error penalty we have in mind is not the usual squared loss: if we predict $\widehat{y}$ and the true value is y, then the penalty should be the absolute difference, $|y - \widehat{y}|$. Write down the loss function that corresponds to the total penalty on the training set.

Logistic regression

4. We identified *inherent uncertainty* as one reason why it might be difficult to get perfect classifiers, even with a lot of training data. In which of the following situations is there likely to be a significant amount of inherent uncertainty?

   (a) x is a picture of an animal and y is the name of the animal

   (b) x consists of the dating profiles of two people and y is whether they will be interested in each other

(c) x is a speech recording and y is the transcription of the speech into words

(d) x is the recording of a new song and y is whether it will be a big hit

5. A logistic regression model given by parameters $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ is fit to a data set of points $x \in \mathbb{R}^d$ with binary labels $y \in \{-1, 1\}$. Consider the set of points x for which the following holds: $\Pr(y = 1|x) = c$.

   (a) For what value of c is this set of points the *decision boundary* of the classifier?

   (b) Now consider the case $c = 3/4$. How is this set of points related to that in (a)? A one-sentence description is sufficient.

   (c) Finally, consider $c = 1/4$. How is this set of points related to those in (a), (b)?

Programming exercises

The data files needed for this week's assignment are in `week4.zip`*, which you should download from the course webpage.*

6. *Discovering relevant features in regression.* The data file `mystery.dat` contains pairs (x, y), where $x \in \mathbb{R}^{100}$ and $y \in \mathbb{R}$. There is one data point per line, with comma-separated values; the very last number in each line is the y-value.

   In this data set, y is a linear function of just *ten* of the features in x, plus some noise. Your job is to identify these ten features.

   (a) Explain your strategy in one or two sentences. Hint: you will find it helpful to look over the routines in `sklearn.linear_model`.

   (b) Which ten features did you identify? You need only give their coordinate numbers, from 1 to 100.

7. *Binary logistic regression.*

   The `heart disease` data set is described at:

   `https://archive.ics.uci.edu/ml/datasets/Heart+Disease`

   The course webpage has a file `heart.csv` that contains a more compact version of this data set with 303 data points, each of which has a 13-dimensional attribute vector x (first 13 columns) and a binary label y (final column). We'll work with this smaller data set.

   (a) Randomly partition the data into 200 training points and 103 test points. Fit a logistic regression model to the training data and display the coefficients of the model. If you had to choose the three features that were most influential in the model, what would they be?

   (b) What is the test error of your model?

   (c) Estimate the error by using 5-fold cross-validation on the training set. How does this compare to the test error?