CSE 151A: Machine learning

Homework 5

Instructions:

- You may discuss problems with your study group, but ultimately all your work (mathematical problems, code, experimental details) must be individual.
- Your solutions must be `typed up` and uploaded to Gradescope by 9.59PM on Thursday November 2. No late homeworks will be accepted under any circumstances, so you are encouraged to upload early.
- A subset of the problems will be graded.

Conceptual and mathematical problems

1. 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
2. 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 $y \in \{-1, 1\}$. Write down a precise expression for the set of points x with
   (a) $\Pr(y = 1|x) = 1/2$
   (b) $\Pr(y = 1|x) = 3/4$
   (c) $\Pr(y = 1|x) = 1/4$
3. *Form of the squashing function.* For $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \{1, 2\}$, consider a distribution over $\mathcal{X} \times \mathcal{Y}$ of the following form:
   - $\Pr(y = 1) = \Pr(y = 2) = 1/2$
   - The distribution of x given $y = 1$ is a spherical Gaussian $N(\mu_1, \sigma^2 I_d)$ and the distribution of x given $y = 2$ is $N(\mu_2, \sigma^2 I_d)$. Recall that the density of $N(\mu, \sigma^2 I_d)$ is given by

$$p(x) = \frac{1}{(2\pi)^{d/2}\sigma^d} \exp\left(-\frac{\|x - \mu\|^2}{2\sigma^2}\right).$$

   Derive a closed-form formula for $\Pr(y = 1|x)$. How does it relate to the squashing function?

4. When using a logistic regression model with two labels, define the *margin* on a point x to be how far its conditional probability is from 1/2:

$$\text{margin}(x) = \left| \Pr(y = 1|x) - \frac{1}{2} \right|.$$

This is a number in the range $[0, 1/2]$.

For any $m \in [0, 1/2]$, define the following two quantities based on a **test set**:

- $f(m)$: the fraction of test points that have margin $\geq m$
- $e(m)$: the error rate on test points with margin $\geq m$

As m grows, how will $f(m)$ and $e(m)$ behave? Would we expect them to increase/decrease? Will they necessarily increase/decrease?

Programming problems

5. *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?

6. *Stepwise forward selection.*

Continuing from the previous problem, suppose we want a **sparse** solution: one that uses only a subset S of the 13 coordinates. One way to do this is with ℓ_1-regularized logistic regression. Another method, which we'll investigate here, is **stepwise forward selection**. This is a greedy procedure that chooses one feature at a time. If we want k features total, these features are selected as follows:

- Let S be empty (this is the set of chosen features)
- Repeat k times:
  - For every feature $f \notin S$:
    - Estimate the error of a classifier based on features $S \cup \{f\}$
  - Select the feature f with the smallest error estimate
  - Add this feature to S
- Now learn a model based only on features S

(a) Use this procedure to find a k-sparse logistic regression solution for the `heart disease` data, for $k = 1, 2, \ldots, 13$. Create a single plot showing the test error and cross-validation error for all these values of k.

(b) What two features were chosen for $k = 2$? Plot the decision boundary in this case.