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 11.59PM on Thursday May 8. 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. *Regression with $d+1$ data points in d dimensions.* In lecture, we asserted that in a regression problem where $\mathcal{X} = \mathbb{R}^d$, it is possible to perfectly fit (almost) any set of d+1 points $(x^{(0)}, y^{(0)}), (x^{(1)}, y^{(1)}), \ldots, (x^{(d)}, y^{(d)})$. Let's see how this works in the specific case where:

   - $x^{(0)} = 0$ (the all-zeros vector)
   - $x^{(i)}$ is the ith coordinate vector (the vector that has a 1 in position i, and zeros everywhere else), for $i = 1, \ldots, d$
   - the response values are $y^{(i)} = c_i$, where $c_0, c_1, \ldots, c_d$ are arbitrary constants.

   Find $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ such that $w \cdot x^{(i)} + b = y^{(i)}$ for all i. You should express your answer in terms of $c_0, c_1, \ldots, c_d$.

2. *Effect of regularization in regression.* Keep the same set of d+1 points $(x^{(0)}, y^{(0)}), (x^{(1)}, y^{(1)}), \ldots, (x^{(d)}, y^{(d)})$ from the previous problem. As we saw, we can find w, b that perfectly fit these points; hence *least-squares regression* would find this "perfect" solution and have zero loss on the training set.

   Now, let us instead use *ridge regression*, with parameter $\lambda \geq 0$, to obtain a solution. We can denote this solution by w_λ, b_λ; notice that it depends upon the choice of λ. Also define the squared training loss associated with this solution,

   $$L_\lambda = \sum_{i=0}^{d} (y^{(i)} - (w_\lambda \cdot x^{(i)} + b_\lambda))^2.$$

   (a) What is L_0 (the training loss when $\lambda = 0$)?
   (b) As λ increases, does $\|w_\lambda\|$ increase, decrease, or stay the same?
   (c) As λ increases, does L_λ increase, decrease, or stay the same?
   (d) As λ goes to infinity, what value does L_λ approach? Your answer should be in terms of the coefficients c_i.

3. 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

4. Consider a classification task with data space $\mathcal{X} = \mathbb{R}^d$, binary labels $\mathcal{Y} = \{-1, +1\}$, and a training set of four points, $(x^{(i)}, y^{(i)})$, $i = 1, 2, 3, 4$. Suppose we are choosing between two different logistic regression models, $w' \cdot x + b'$ and $w'' \cdot x + b''$, which behave as follow on the training data:

Index i	$y^{(i)}$	$w' \cdot x^{(i)} + b'$	$w'' \cdot x^{(i)} + b''$
1	$+1$	2.3	3.1
2	-1	0.2	9.2
3	-1	-1.1	-0.1
4	$+1$	-0.1	-1.0

(a) Suppose these two models are used to predict labels for the four data points. How many mistakes does model (w', b') make? How many mistakes does (w'', b'') make?

(b) Recall that the logistic loss of a model (w, b) on a data set $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$ is given by

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

What is the logistic loss of (w', b') on the data set of four points? What is the loss of (w'', b'')?

(c) How would you account for the discrepancy between parts (a) and (b)?

5. 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 $[0, 1/2]$. (We can think of this as signifying how confident the model is in its prediction: the larger the margin, the more confident.)

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 $\geq m$

(a) As m grows, will $f(m)$ increase or decrease?

(b) As m grows, would be expect $e(m)$ to increase or decrease? Will it necessarily behave in this way?

Programming problems

6. *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.

   (b) If you had to choose the three features that were most influential in the model, what would they be? Explain the basis for your selection.

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

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

7. *Stepwise forward selection.* For this problem, we will use the same `heart.csv` data set as in the previous problem.

   Now 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) Implement this stepwise forward selection algorithm. You might find it helpful to write a function `ErrorEstimate`(x, y, S) which:
   - takes as input a data set (x, y) and a list of features S,
   - fits a logistic regression model to the data, restricted to the selected features, using `sklearn` (and making sure not to regularize, i.e. `penalty = None`),
   - and estimates the error of this classifier using cross-validation.

   (b) As you did in the previous problem, load in the `heart` data set and split it into a training set and test set. Use your stepwise forward selection procedure to fit a k-sparse logistic regression model to the training data, for all values $k = 1, 2, \ldots, 13$. Create a single plot showing the test error and cross-validation error for all these values of k.

   (c) What two features were chosen for $k = 2$? Plot the decision boundary in this case (in terms of just these two features).