CS155 Homework 1

Ty Limpasuvan

4 late hours used, 44 remaining

1 Basics

1.1 A

A hypothesis set is the set of possible hypotheses that potentially model an unknown target function.

1.2 B

The hypothesis set of a linear model is the set of lines that could serve the same function as the target.

1.3 C

Overfitting refers to the act of fitting too closely to the given data. This often results from fitting to noise, and makes the estimation have large amounts of out of sample error.

1.4 D

One way to help prevent overfitting is to decrease the amount of noise. Another way is to increase the number of data points.

1.5 E

Training data is the data we use to create our hypothesis used for estimations, and test data is the data we attempt to use our hypothesis on for classifications. The test data is used to see how well the estimate performs, but the test data was not intended for the estimate to learn from it.

1.6 F

We assume that the points in the dataset are sample randomly and that the sets used for training and testing are sampled from the same distribution.

1.7 G

The input space could be the English language and the output space could be +1 and -1 to signify whether the email is spam.

1.8 H

It is the act of breaking data into k folds, and run multiple training sessions with one fold for validation and the others for training.

2 Bias-Variance Tradeoff

2.1 A

E$_s[(f_s(x) - f(x))^2] = E_s[f_s(x) - F(x))^2] + (F(x) - f(x))^2$
$E_s[E_x(f_s)] = E_x[E_s[(f_s(x) - f(x))^2]]$
$= E_x[Bias(x) + Var(x)]$

2.2 C

Degree 1 has the highest bias. As N increases, the out of sample error (validation) should approach the bias. Degree 1 has the greatest validation error at N = 100 out of the different graphs.

2.3 D

Degree 12 has the highest variance. The curve above the convergence toward bias represents the variance. At small N values, degree 12 has the greatest validation error.

2.4 E

We observe the out of sample error decrease with increasing N values. We also begin to see some convergence with the higher values of N.

2.5 F

The model we use has already considered the points that were used for training. So when comparing in and out of sample error, it is expected that out of sample error is greater. The possibility of overfitting to the training data also exists, which would decrease the in sample error and increase the out of sample error.

2.6 G

Degree 6 seems that it would perform the best in this situation; both its in and out of sample errors were the lowest at high values of N.

3 The Perceptron

3.1 B

In a 2D dataset, 4 data points may not be linearly separable. For example, a point classified as +1 could be inside of a triangle of 3 points classified as -1. A single line cannot properly separate the two groups. In a 3D dataset, 5 points are needed at minimum. For example, consider the case of having three points classified as -1 in a plane and two points classified as +1 on opposite sides of the plane. A single plane is unable to properly separate these classified points correctly. For the N-dimensional set, no ¡N-dimensional hyperplane contains a non-linearly-separable subset if there are N + 2 points.

3.2 C

The PLA will never converge. We see that the process enters a cycle after the first update. The algorithm will keep repeating by updating the weight vector on a single misclassified point. And because the dataset is not linearly separable, the weight vector will always be updated and the PLA will never reach the point where all of the points are correctly classified by the line.

4 Stochastic Gradient Descent

4.1 A

We would prepend a 1 onto the weight vector and prepend the b term to the x vector.

4.2 B

$\partial_w (y - f(x))^2$
$\partial_w (y - w^T x)^2$
$-2(y - w^T x) * \partial_w (y - w^T x)$
$-2(y - w^T x)x$

4.3 D

Regardless of the start point, all of the paths travel down the gradient and converge to the same point. The paths behave the same way in the different datasets. They still travel down the gradient and converge to the same point.

4.3.1 E

The larger the eta, the slower the loss value approached a certain value. The graphs with smaller eta values approached this value much faster.

4.4 G

The greater the eta value, the faster the plot converged, just like before.

4.5 H

My result doesn't match very well with my result from SGD.

4.6 I

If the data is taken in one point at a time, SGD only needs to have the gradient updated, while a closed form solution requires that all of the iteration be done.

4.7 J

Whenever the different eta graphs are acceptably close to one another (in convergence), the process could be stopped.

4.8 K

For a perceptron, the convergence is found when all of the points have been classified correctly. But for SGD, the weight vector is continuously adjusted until either a time limit is hit or a minimum is found.