CIS 519 Problem Set 2

Gabrielle Merritt

10/13/2014

Part I
Problem Set

1 Gradient Descent

Let k be a counter for the iterations of gradient descent, and let α_k be the learning rate for the k^{th} step of gradient descent. In one sentence, what are the implications of using a constant value for α_k in gradient descent? In another sentence, what are the implications for setting α_k as a function of k?

Using α_k as a constant ensures that your step size along the gradient is constant; therefore, the number of iterations doesn't affect the speed at which you arrive at the minima. If α_k is a function of k then the size of your step changes proportionally to k, and you may arrive at the minima much faster (or slower)

2 Fitting an SVM by Hand

Consider a dataset with only 2 points in 1D: $(x_1 = 0, y_1 = -1)$ and $(x_2 = \sqrt{2}, y_2 = +1)$. Consider mapping each point to 3D using the feature vector $\phi(x) = [1,\sqrt{2}x, x^2]^T$ (i.e., use a 2nd order polynomial kernel). The maximum margin classifier has the form

$$min||w||_2^2 s.t.(1)$$

$$y_1(w^T\phi(x_1) + w_0) \geq 1(2)$$

$$y_2(w^T\phi(x_2) + w_0) \geq 1(3)$$

2.1 vector parallel to optimal vector w

Solution: Any vector that isn't linearly independent of w is parallel, so if c is a constant then we can say that

$$c * w||w$$

in order to obtain $min||w||_2^2$ the optimal vector w is

$$w = [.5, 0, 0]$$

a possible parallel vector

$$2 * w = [1, 0, 0]$$

2.2 What is the value of the margin that is achieved by w?

Solution: The margin is defined as two times the optimal vector

$$\frac{2}{||w||_2}$$

2.3 Solve for w, using the fact that the margin is equal to $\frac{2}{||w||_2}$

Solution:

$$\frac{2}{||w||_2} = 4$$

2.4 Solve for w_0, using your value of w and the two constraints (2)-(3) for the max margin classifier.

Solution:

$$w = [w_0, w_1, w_2]$$
$$\phi(x_1) = [1, 0, 0]^T$$
$$\phi(x_2) = [1, 2, 2]^T$$
$$y_1(w^T\phi(x_1) + w_0) \geq 1 = -1(2 * w_0) \geq 1$$
$$w_0 \leq .5$$
$$y_2(w^T\phi(x_2) + w_0) \geq 1 = (2w_0 + 2w_1 + 2w_2) \geq 1$$
$$-(w_1 + w_2) \leq w_0$$

the smallest norm you can find for vector w is with w_1 and $w_2 = 0$

$$w = [.5, 0, 0]$$

2.5 Write down the form of the discriminant $h(x) = w_0 + w^T\phi(x)$

Solution: Discriminate for second order polynomial $ax^2 + bx + c$ is notated as $b - 4ac$ for our problem it can be re written as

$$h(x) = w_2x^2 + \sqrt{2}w_1x + w_0$$
$$D = 2 * w_1^2 - 4 * w_2 * w_0$$

3 Support Vectors

For an SVM, if we remove one of the support vectors from the training set, does the size of the maximum margin decrease, stay the same, or increase for that dataset? Why? (Explain your answer in 1-2 sentences.)

The size of the margin either stays the same or increases since the support vectors are defined as being a set of data points with closest distance to the hyper plane. If a data point that is also a support vector is removed then usually the learner will try to increase the margin to the next closest point in the data set.

4 VC Dimension

4.1

Imagine that we are working in Rd space and we are using a hyper-dimensional sphere centered at the origin as a classifier. Anything inside the sphere is considered positive, and the only thing we can do to train the model is to adjust the radius of the sphere (it stays centred at the origin). What is the VC dimension of this classifier?

Since a sphere is in R^3 space and the VC dimension of a linear classifier is defined as 1+d. The VC dimension of the classifier is 4

4.2

Now, imagine we are able to change the direction of the classification surface, so that we could have anything inside the sphere be predicted positive or everything inside the sphere be predicted negative (our choice). What is the VC dimension of this classifier now?

The VC dimension is now 1, since it cannot shatter two points of different signs within the sphere

5 Generalizing to Unseen Data

I decided to use an SVM with a gaussian kernel because SVMs are good at finding general models for data. I did a search for optimal parameters and ended up using 10 for C and 1 for sigma. I trained the model using the training data provided and training on half, and testing on the other half. Then randomizing the data and repeating the process. Since my test accuracy was consistently above 95 percent I tested it on the unlabelled data.