CS 189 Introduction to Machine Learning
Fall 2017 DIS9

1 Randomized Kaczmarz SGD

In this problem we look at a variant of SGD known as the randomized Kaczmarz method in linear least squares. It may be instructive to first do this exercise for vanilla SGD instead of this variant of it.

The least squares problem can be written as:

$$\min_w \frac{1}{2}||Aw-y||_2^2 = \min_w \frac{1}{2}\Sigma_{i=1}^n (A_i^\top w - y_i)^2 = \min_w \frac{1}{2}\sum_{i=1}^n \ell_i(w)$$

.

Assume that $Aw^* = y$, or in other words, that the minimum loss is 0. Here $A_i^\top$ is the ith row of matrix A and ℓ_i is the loss of the training example i. We implement a variant of SGD as follows:

$$w^t = w^{t-1} - \alpha_J \nabla f_J(w^{t-1})$$

with (distinct) learning rates $\alpha_j = \frac{1}{||A_j||_2^2}$ for each $j = 1,2,\ldots,n$. Above, J is a random index between 1 and n such that the probability of choosing $J = j$ is given by $p(J=j) = \frac{||A_j||_2^2}{||A||_F^2}$ for $j \in \{1,..,n\}$

(a) We define a (random) projection operator $P_J = \alpha_J A_J A_J^\top$. Show that

$$E_J[P_J] = \sum_{i=1}^n p(J=i)P_i = \frac{A^T A}{\|A\|_F^2}$$

.

Solution:

$$\mathbb{E}[P_J] = \sum_{i=1}^n p(J=i)\alpha_i P_i = \sum_{i=1}^n \frac{\|A_i\|_2^2 A_i A_i^\top}{\|A\|_F^2 \|A_i\|_2^2} = \Sigma_{i=1}^n \frac{A_i A_i^\top}{\|A\|_F^2} = \frac{A^T A}{\|A\|_F^2}.$$

(b) Show that stochastic gradient descent will converge according to the following:

$$\mathbb{E}[\left\|w^t - w^*\right\|_2^2] \leq \left(1 - \frac{\sigma_{min}(A)^2}{\|A\|_F^2}\right)\mathbb{E}[\left\|w^{t-1} - w^*\right\|_2^2]$$

.

Solution:

$$\nabla f_j(w^t) = A_j(A_j^\top w^t - y_j) = A_j A_j^\top (w^t - w^*),$$

where we have used the fact that $A_j^\top w^* = y_j$.

Subtract w^* from both sides of the SGD update rule:

$$\begin{aligned} w^t - w^* &= w^{t-1} - w^* - \alpha_j A_j A_j^\top (w^{t-1} - w^*) \\ &= (I - \frac{A_j A_j^\top}{\|A_j\|_2^2})(w^{t-1} - w^*) \end{aligned}$$

Notice that $\frac{A_j A_j^\top}{\|A_j\|_2^2}$ is a projection matrix. Take ℓ_2 norm squared from the above equation to obtain:

$$\begin{aligned} \|w^t - w^*\|_2^2 &= (w^{t-1} - w^*)^T (I - Q_j)^2 (w^{t-1} - w^*) \\ &= \|(w^{t-1} - w^*)\|_2^2 - (w^{t-1} - w^*)^T Q_j (w^{t-1} - w^*). \end{aligned}$$

Here, we have used the shorthand $Q_j = \frac{A_j A_j^\top}{\|A_j\|_2^2}$. In the second line, we notice that since Q_j is a projection matrix, we have $Q^2 = Q$.

We now have an expression for each w^{t-1} and j, but these two are in fact random. We first take the expectation conditioned on w^{t-1} (notice that we are using the fact that the index J is now independent of the randomness up to time $t-1$.)

$$E[\|w^t - w^*\|_2^2 | w^{t-1}] = \|(w^{t-1} - w^*)\|_2^2 - (w^{t-1} - w^*)^T \mathbb{E}[Q_J](w^{t-1} - w^*).$$

Using the result of part (b), we have:

$$E[\|w^t - w^*\|_2^2 | w^{t-1}] = \|(w^{t-1} - w^*)\|_2^2 - (\frac{1}{\|A\|_F^2}(A^T A(w^{t-1} - w^*)))^T ((w^{t-1} - w^*))$$

$$\leq \|(w^{t-1} - w^*)\|_2^2 - \frac{\sigma_{min}(A)^2}{\|A\|_F^2} \|(w^{t-1} - w^*)\|_2^2$$

$$E[\|w^t - w^*\|_2^2 | w^{t-1}] \leq (1 - \frac{\sigma_{min}(A)^2}{\|A\|_F^2})(\|w^{t-1} - w^*\|_2^2)$$

Remember the conditional expectation property $E[E[X|Y]] = E[X]$. So we have $E[E[\|w^t - w^*\|_2^2 | w^{t-1}]] = E[\|w^t - w^*\|_2^2]$. So if we take the expectation from the equation above we get:

$$E[\|w^t - w^*\|_2^2] \leq (1 - \frac{\sigma_{min}(A)^2}{\|A\|_F^2}) E[\|w^{t-1} - w^*\|_2^2]$$

.

(c) What does the result of the previous part mean? Contrast it with the rate of convergence of SGD that you saw in class.

Solution: In this problem we showed that even though we are taking samples from the dataset, stochastic gradient descent mean error converges exponentially fast. Note that here we didn't compute the variance of error and it can be large.

2 Quadratic Discriminant Analysis (QDA)

We have training data for a two class classification problem as laid out in Figure 1. The black dots are examples of the positive class ($y = +1$) and the white dots examples of the negative class ($y = -1$).

(a) Draw on Figure 1 the position of the class centroids $\mu_{(+)}$ and $\mu_{(-)}$ for the positive and negative class respectively, and indicate them as circled $(+)$ and $(-)$. Give their coordinates:

$$\mu_{(+)} = \begin{bmatrix} & \end{bmatrix} \quad \mu_{(-)} = \begin{bmatrix} & \end{bmatrix}$$

Solution:

$$\mu_{(+)} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$$

$$\mu_{(-)} = \begin{bmatrix} 3 \\ 0 \end{bmatrix}$$

(b) Assume each class has data distributed according to a bi-variate Gaussian, centered on the class centroids computed in question (a). Draw on Figure 1 the contour of equal likelihood $p(X = x|Y = y)$ going through the data samples, for each class. Indicate with light lines the principal axes of the data distribution for each class.

(c) Compute the covariance matrices for each class:

$$\Sigma_{(+)} = \begin{bmatrix} & \\ & \end{bmatrix} \quad \Sigma_{(-)} = \begin{bmatrix} & \\ & \end{bmatrix}$$

Solution:

$$\Sigma_{(+)} = \begin{bmatrix} 1/2 & 0 \\ 0 & 2 \end{bmatrix}$$

$$\Sigma_{(-)} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

(d) Compute the determinant and the inverse of $\Sigma_{(+)}$ and $\Sigma_{(-)}$:

$$|\Sigma_{(+)}| = \qquad\qquad |\Sigma_{(-)}| =$$

$$\Sigma_{(+)}^{-1} = \begin{bmatrix} \quad & \quad \\ \quad & \quad \end{bmatrix} \qquad\qquad \Sigma_{(-)}^{-1} = \begin{bmatrix} \quad & \quad \\ \quad & \quad \end{bmatrix}$$

Solution:

$$|\Sigma_{(+)}| = 1, |\Sigma_{(-)}| = 4$$

$$\Sigma_{(+)}^{-1} = \begin{bmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$$

$$\Sigma_{(-)}^{-1} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$$

(e) The likelihood of examples of the positive class is given by:

$$p(X = x|Y = +1) = \frac{1}{2\pi|\Sigma_{(+)}|^{1/2}} \exp\big(-\frac{1}{2}(x-\mu_{(+)})^T\Sigma_{(+)}^{-1}(x-\mu_{(+)})\big)$$

and there is a similar formula for $p(X = x|Y = -1)$. Compute $f_{(+)}(x) = \log\big(p(X = x|Y = +1)\big)$ and $f_{(-)}(x) = \log\big(p(X = x|Y = -1)\big)$. Then compute the discriminant function $f(x) = f_{(+)}(x) - f_{(-)}(x)$:

$$f_{(+)}(x) =$$

$$f_{(-)}(x) =$$

$$f(x) =$$

Solution:

$$f_{(+)}(x) = -\frac{1}{2}(x-\mu_{(+)})^T\Sigma_{(+)}^{-1}(x-\mu_{(+)}) - \log(2\pi|\Sigma_{(+)}|)$$
$$= -(x_1-3)^2 - \frac{1}{4}(x_2-3)^2 - \log(2\pi)$$

$$f_{(-)}(x) = -\frac{1}{2}(x-\mu_{(-)})^T\Sigma_{(-)}^{-1}(x-\mu_{(-)}) - \log(2\pi|\Sigma_{(-)}|^{1/2})$$
$$= -\frac{1}{4}(x_1-3)^2 - \frac{1}{4}x_2^2 - \log(4\pi)$$

$$f(x) = -\frac{1}{2}((x-\mu_{(+)})^T\Sigma_{(+)}^{-1}(x-\mu_{(+)}) - (x-\mu_{(-)})^T\Sigma_{(-)}^{-1}(x-\mu_{(-)})) - \log(\frac{|\Sigma_{(+)}|}{|\Sigma_{(-)}|})$$
$$= -\frac{3}{4}(x_1-3)^2 + \frac{3}{2}x_2 - \frac{9}{4} + \log(2)$$

(f) Draw on Figure 1 for each class contours increasing equal likelihood. Geometrically construct the Bayes optimal decision boundary. Compare to the formula obtained with $f(x) = 0$ after expressing x_2 as a function of x_1:

$x_2 =$

What type of function is it?

Solution: We put $f(x) = 0$ so:

$$x_2 = \frac{1}{2}(x_1-3)^2 + \frac{3}{2} - \frac{2}{3}\log(2)$$

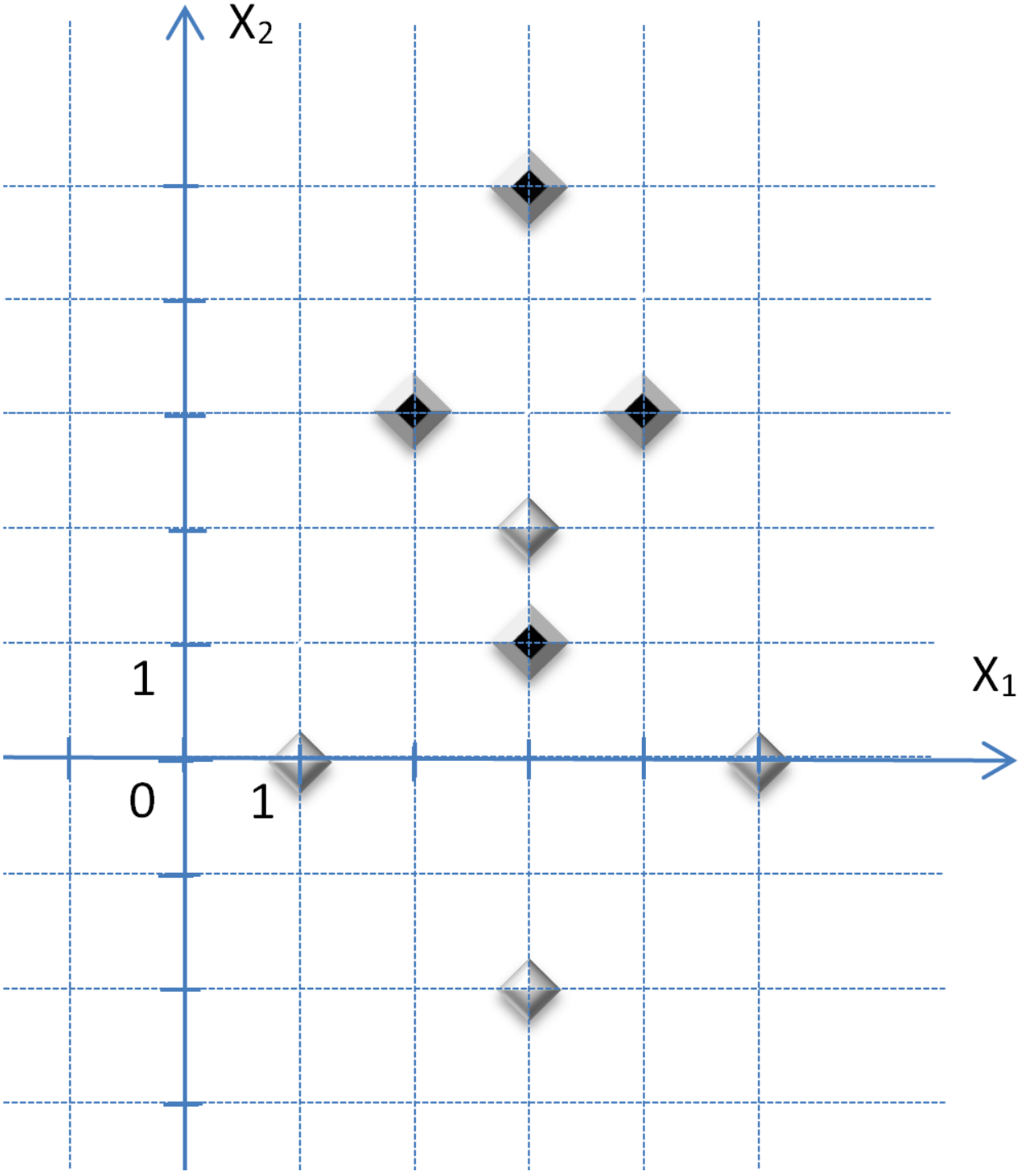

Figure 1: Draw your answers to the QDA problem.