1 Simple SGD updates

Let us consider a simple least squares problem, where we are interested in optimizing the function

$$F(w) = \frac{1}{2n}\|Aw - y\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}(a_i^\top w - y_i)^2.$$

1. What is the closed form OLS solution? **What is the time complexity of computing this solution in terms of flops?**

2. Write down the SGD update for logistic regression on two classes

$$F(w) = \frac{1}{n}\sum_{i=1}^{n} y_i \log\frac{1}{\sigma(w^\top x_i)} + (1 - y_i)\log\frac{1}{1 - \sigma(w^\top x_i)}.$$

   Discuss why this is equivalent to minimizing a "cross-entropy" loss.

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 ($y = -1$).

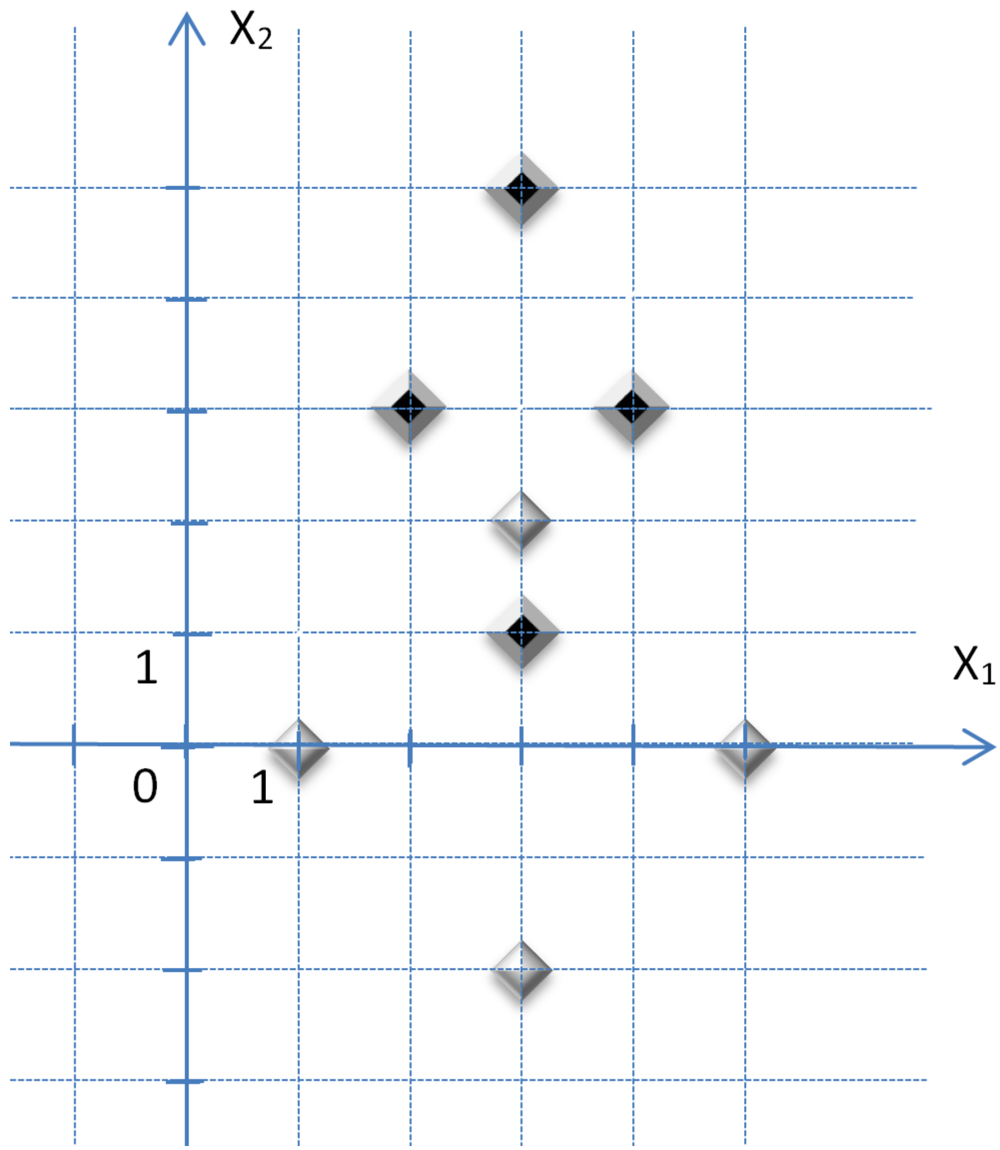


Figure 1: Draw your answers to the QDA problem.

1. 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:

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

2. Compute the covariance matrices for each class:

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

3. 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.

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

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

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

5. 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 $f(x) = f_{(+)}(x) - f_{(-)}(x)$:

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

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

$f(x) =$

6. 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:

$x_2 =$

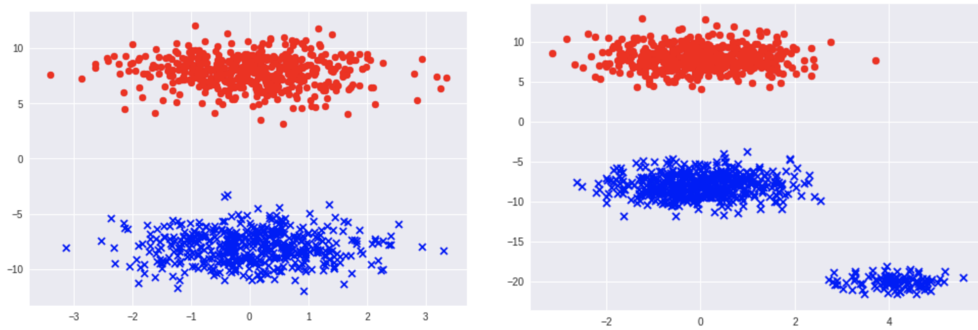
What type of function is it?

7. Now assume $p(Y = -1) \neq p(Y = +1)$, how does it change the decision boundary?

3 Logistic Regression

In this problem, we will explore logistic regression and derive some insights.

1. You are given the following datasets:



Assume you are using Least Square Means for classification. Draw the decision boundary for the dataset above. Recall that the optimization problem has the following form:

$$\arg\min_{\mathbf{w}} \sum_{i=1}^{n} (\mathbf{w}^\top \mathbf{x}_i - y_i)^2 + \lambda \|\mathbf{w}\|_2^2$$

2. Draw the ideal decision boundary for the dataset above.

3. Assume your data comes from two classes and the prior for class k is $p(y = k) = \pi_k$. Also the conditional probability distribution for each class k is Gaussian, $\mathbf{x}|y = k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, that is $f_k(\mathbf{x}) = f(\mathbf{x}|y = k) = \frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}|}} \exp\left\{(\mathbf{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_k)\right\}$. Assume that $\{\boldsymbol{\mu}_k\}_{k=0}^{1}, \boldsymbol{\Sigma}$ where estimated from the training data.

   Show that $P(y|\mathbf{x}) = s(\mathbf{w}^\top \mathbf{x})$ is the sigmoid function, where $s(\zeta) = \frac{1}{1+e^{-\zeta}}$.

4. In the previous part we saw that the posterior probability for each class is the sigmoid function under the LDA model assumptions. Notice that LDA is a generative model. In this part we are going to look at the discriminative model. We will assume that the posterior probability has Bernoulli distribution and the probability for each class is the sigmoid function, i.e. $p(y|\mathbf{x}; \mathbf{w}) = q^y(1 - q)^{1-y}$, where $q = s(\mathbf{w}^\top \mathbf{x})$ and try to find $\mathbf{w}$ that maximizes the likelihood function. **Can you find a closed form maximum-likelihood estimation of $\mathbf{w}$?**

5. In this section we are going to use Newton method to find the optimal solution for $\mathbf{w}$. **Write out the update step of Newton method. What other method does this resemble?**

