CS 189 Introduction to Machine Learning
Spring 2018 **DIS9**

1 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 ($y = -1$).

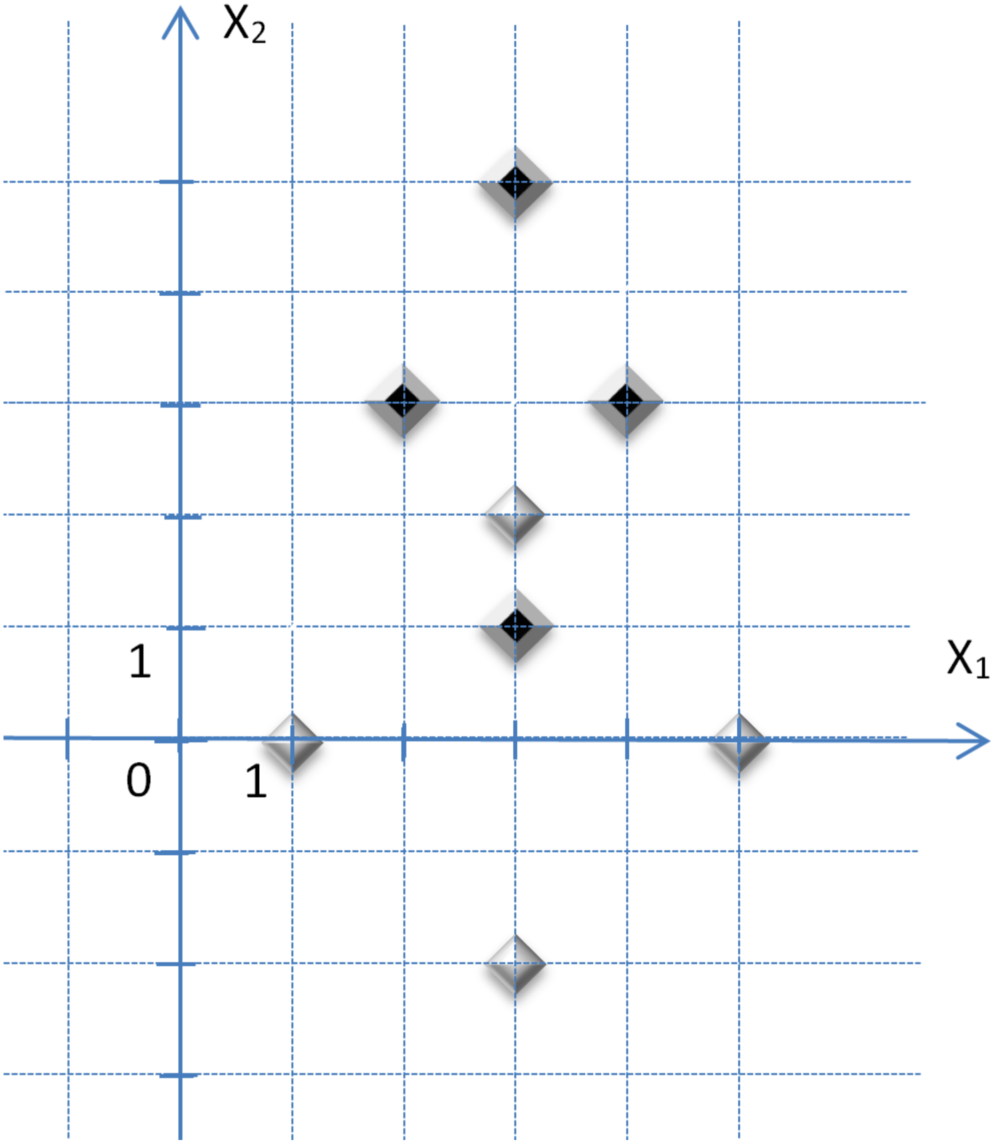


Figure 1: Draw your answers to the QDA problem.

(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) Compute the covariance matrices for each class:

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

Solution:

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

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

(c) 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.

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

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

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

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:

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

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

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

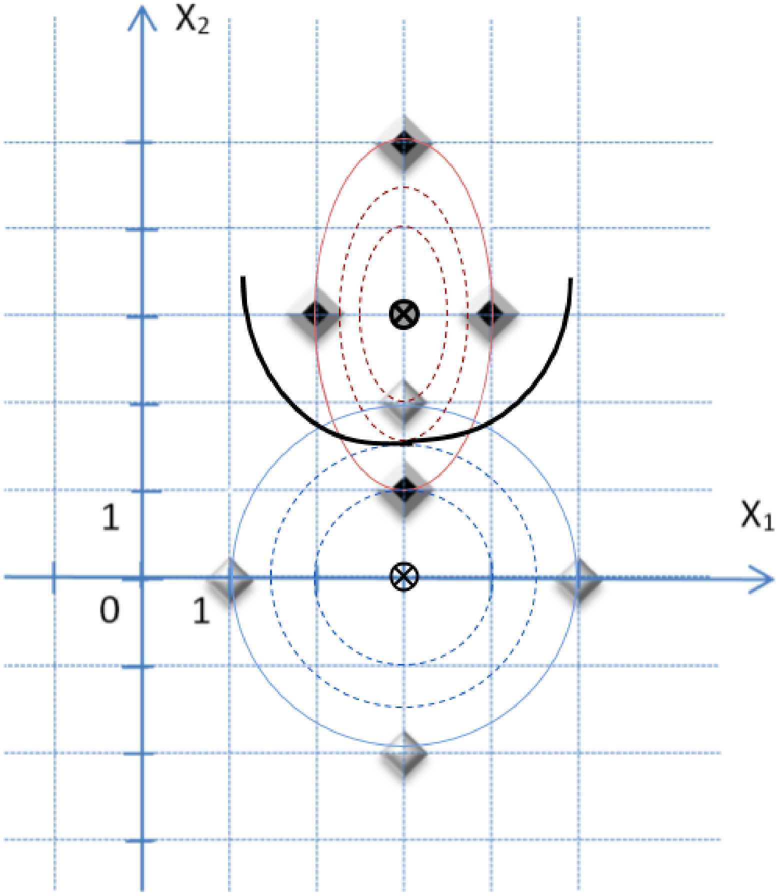
(f) Draw on Figure 1 for each class contours increasing equal likelihood. Geometrically construct the Bayes optimal decision boundary. Compare to the formula obtained with $f(x) = 0$ after expressing x_2 as a function of x_1:

$x_2 =$

What type of function is it?

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

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



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

Solution: We get that:

$$\log\big(p(X = x, Y = +1)\big) = f_{(+)}(x) + \log\big(p(Y = +1)\big)$$

$$\log\big(p(X=x,Y=-1)\big) = f_{(-)}(x) + \log\big(p(Y=-1)\big)$$

To find the boundary we need to find where $\log\big(p(X=x,Y=+1)\big) = \log\big(p(X=x,Y=-1)\big)$:

$$f_{(+)}(x) + \log\big(p(Y=+1)\big) = f_{(-)}(x) + \log\big(p(Y=-1)\big)$$

$$f(x) + \log\left(\frac{p(Y=+1)}{p(Y=-1)}\right) = 0$$

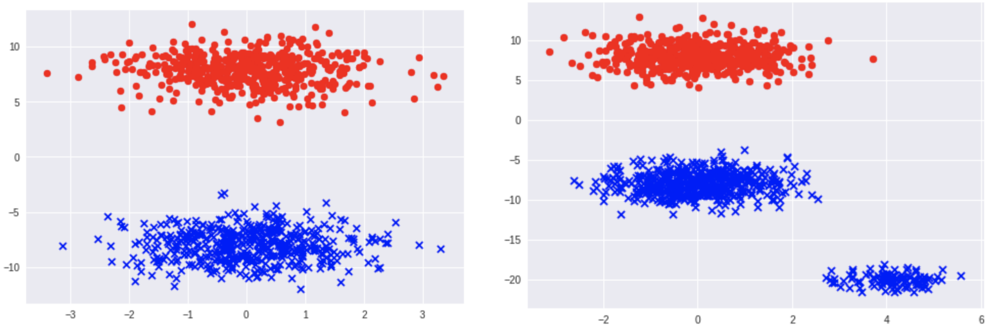
$$x_2 = \frac{1}{2}(x_1-3)^2 + \frac{3}{2} - \frac{2}{3}\log(2) + \frac{2}{3}\log\frac{p(Y=-1)}{p(Y=+1)}$$

The boundary is shifted by $\frac{2}{3}\log\frac{p(Y=-1)}{p(Y=+1)}$

2 Logistic Regression

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

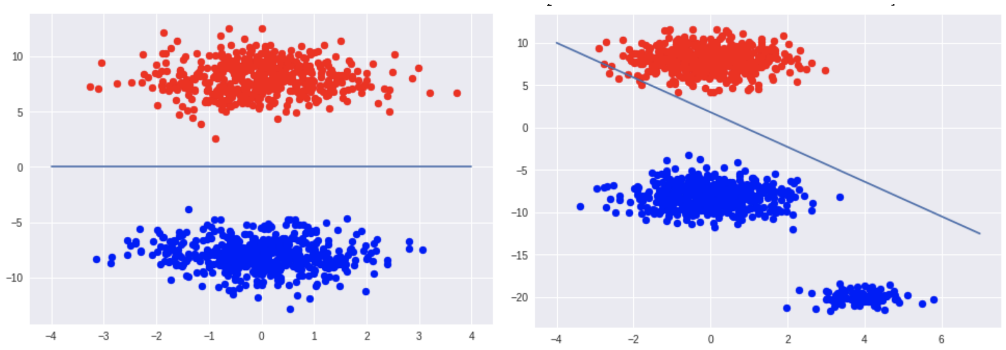
(a) You are given the following datasets:



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:

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

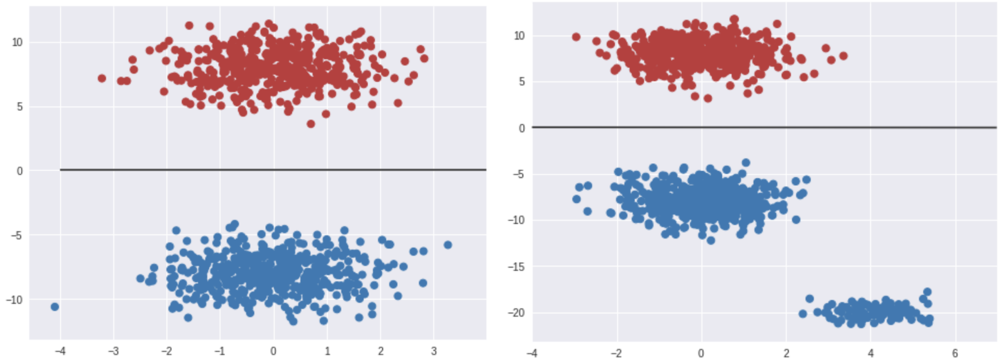
Solution:



During the optimization process, the magnitude of $\mathbf{w}^\top \mathbf{x}$ is used, but we will classify a point based on $\operatorname{sign}(\mathbf{w}^\top \mathbf{x})$. Notice that even though the decision boundary on the first dataset would be valid for the rest, because there is a cluster of points in the bottom right corner the magnitude of the error would be higher for those points, pulling the decision boundary down.

(b) Draw the ideal decision boundary for the dataset above.

Solution:



Ideally we wouldn't want the decision boundary to change once we add the points in the right bottom corner.

(c) 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 $f_k(\mathbf{x}) = f(\mathbf{x}|y=k) = \frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}|}} \exp\Big\{(\mathbf{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\Big\}$. 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 $s(\zeta) = \frac{1}{1+e^{-\zeta}}$.

Solution: Let us denote $Q_k(\mathbf{x}) = \log\Big((\sqrt{2\pi})^d \pi_k f_k(\mathbf{x})\Big)$, so we get that

$$\begin{aligned}
p(y=1|\mathbf{x}) &= \frac{\pi_1 f_1(\mathbf{x})}{\pi_0 f_0(\mathbf{x}) + \pi_1 f_1(\mathbf{x})} = \left(1 + \frac{\pi_0 f_0(\mathbf{x})}{\pi_1 f_1(\mathbf{x})}\right)^{-1} \\
&= \left(1 + \frac{e^{Q_0(\mathbf{x})}}{e^{Q_1(\mathbf{x})}}\right)^{-1} = s(Q_1(\mathbf{x}) - Q_0(\mathbf{x}))
\end{aligned}$$

Now lets look at the expression $Q_1(\mathbf{x}) - Q_0(\mathbf{x})$

$$\begin{aligned}
Q_1(\mathbf{x}) - Q_0(\mathbf{x}) &= \log\Big((\sqrt{2\pi})^d \pi_1 f_1(\mathbf{x})\Big) - \log\Big((\sqrt{2\pi})^d \pi_0 f_0(\mathbf{x})\Big) \\
&= \log\frac{\pi_1}{1-\pi_1} - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_0) \\
&= \log\frac{\pi_1}{1-\pi_1} + \mathbf{x}^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) - \frac{1}{2}\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_0^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0
\end{aligned}$$

Notice that we can write it out as:

$$Q_1(\mathbf{x}) - Q_0(\mathbf{x}) = \log \frac{\pi_1}{1-\pi_1} + \frac{1}{2}\boldsymbol{\mu}_0^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0 - \frac{1}{2}\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} = w_0 + \mathbf{w}^\top \mathbf{x}$$

(d) 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 w?**

Solution:

Assume that our dataset is of size n. So we get that the likelihood is:

$$L(\mathbf{w}) = \prod_{i=1}^n p(y = y_i|\mathbf{x}_i) = \prod_{i=1}^n q_i^{y_i}(1-q_i)^{1-y_i}.$$

Now maximizing the likelihood of the training data as a function of the parameters $\mathbf{w}$, we get:

$$\begin{aligned}\hat{\mathbf{w}} &= \arg\max_{\mathbf{w}} \ L(\mathbf{w}) = \arg\max_{\mathbf{w}} \prod_{i=1}^n q_i^{y_i}(1-q_i)^{1-y_i} \\ &= \arg\max_{\mathbf{w}} \sum_{i=1}^n y_i \log(q_i) + (1-y_i)\log(1-q_i) \\ &= \arg\max_{\mathbf{w}} \sum_{i=1}^n y_i \log\left(\frac{q_i}{1-q_i}\right) + \log(1-q_i)\end{aligned}$$

Since q_i is the sigmoid function, we get that:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} - \sum_{i=1}^n y_i \mathbf{w}^\top \mathbf{x}_i - \log\left(1 + \exp\left\{\mathbf{w}^\top \mathbf{x}_i\right\}\right)$$

.

Let us denote $J(\mathbf{w}) = -\sum_{i=1}^n y_i \mathbf{w}^\top \mathbf{x}_i - \log\left(1 + \exp\left\{\mathbf{w}^\top \mathbf{x}_i\right\}\right)$. Notice that $J(\mathbf{w})$ is convex in $\mathbf{w}$, so global minima can be found. Recall that $s'(\zeta) = s(\zeta)(1-s(\zeta))$. Now let us take the derivative of $J(\mathbf{w})$ w.r.t $\mathbf{w}$:

$$\frac{\partial J}{\partial \mathbf{w}} = -\sum_{i=1}^n y_i \mathbf{x}_i - \frac{\exp\left\{\mathbf{w}^\top \mathbf{x}_i\right\}}{1 + \exp\left\{\mathbf{w}^\top \mathbf{x}_i\right\}}\mathbf{x}_i = \sum_{i=1}^n (s(\mathbf{w}^\top \mathbf{x}_i) - y_i)\mathbf{x}_i = \sum_{i=1}^n (s_i - y_i)\mathbf{x}_i = \mathbf{X}^\top(\mathbf{s} - \mathbf{y})$$

where, $s_i = s(\mathbf{w}^\top \mathbf{x}_i), \mathbf{s} = (s_1, \ldots, s_n)^\top, \mathbf{y} = (y_1, \ldots, y_n)^\top$ and $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_n^\top \end{bmatrix}$

Can't get a closed form estimate for $\mathbf{w}$ by setting the derivative to zero.

(e) In this section we are going to use Newton method to find the optimal solution for w. **Write out the update step of Newton method. What other method does this resemble?**

Solution: In the previous section we saw that we couldn't find a closed form solution for $\hat{\mathrm{w}}$, so to solve this problem we are going to use Newton method. Newton method, is an iterative method for finding successively better approximations to the roots (or zeroes) of a real-valued function. The iterative k step in Newton method for some function $f(\mathbf{x})$ is:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - (\nabla f(\mathbf{x}^{(k)}))^{-1} f(\mathbf{x}^{(k)})$$

In our case we want to find the zeros of $\nabla J(\mathbf{w})$. The update step will be:

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - (HJ(\mathbf{w}^{(k)}))^{-1} \nabla_w J(\mathbf{w}^{(k)})$$

Where,

$$\nabla_w J(\mathbf{w}) = \mathbf{X}^\top (\mathbf{s} - \mathbf{y})$$
$$HJ(\mathbf{w}) = \nabla_w^2 J(\mathbf{w}) = \nabla_w \mathbf{X}^\top (\mathbf{s} - \mathbf{y}) = \sum_{i=1}^{n} s_i(1 - s_i)\mathbf{x}_i \mathbf{x}_i^\top = \mathbf{X}^\top \mathbf{\Omega} \mathbf{X}$$

where, $\mathbf{\Omega} = \mathrm{diag}(s_1(1 - s_1), \cdots, s_n(1 - s_n))$

Finding the inverse of the Hessian in high dimensions can be an expensive operation. Instead of directly inverting the Hessian we calculate the vector $\Delta_{k+1} = \mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}$ as the solution to the system of linear equations:

$$HJ(\mathbf{w}^{(k)})\Delta_{k+1} = -\nabla J(\mathbf{w}^{(k)})$$
$$\mathbf{X}^\top \mathbf{\Omega} \mathbf{X} \Delta_{k+1} = \mathbf{X}^\top (\mathbf{y} - \mathbf{s})$$

This looks similar to weighted least squares where $\mathbf{\Omega}$ and $(\mathbf{y} - \mathbf{s})$ change per iteration. Specifically looking at points where $s(\mathbf{w}^\top \mathbf{x}_i)$ is close to 0.5, Ω_i has the highest weight and points where $s(\mathbf{w}^\top \mathbf{x}_i)$ is close to 0 or 1 has much lower weight. Points where $s(\mathbf{w}^\top \mathbf{x}_i) \approx 0.5$ are close to the decision boundary and model is least sure of these points' cluster, so they move the decision boundary most in the next iteration.