Linear classification

Q1:

a)

The posterior distribution is a sigmoid of a linear function, or equivalently the Bernoulli distribution.

b)

According to the Bayes' Rule,

$$\begin{aligned} p(y=1|x) =& \frac{p(x|y=1)p(y=1)}{p(x|y=0)p(y=0)+p(x|y=1)p(y=1)} \\ p(y=0|x) =& \frac{p(x|y=0)p(y=0)}{p(x|y=0)p(y=0)+p(x|y=1)p(y=1)} \end{aligned} \tag{1}$$

As

$$p(y=0) = p(y=1) = \frac{1}{2}$$

then

$$\begin{aligned} p(y=1|x) =& \frac{p(x|y=1)}{p(x|y=0)+p(x|y=1)} \\ p(y=0|x) =& \frac{p(x|y=0)}{p(x|y=0)+p(x|y=1)} \end{aligned} \tag{2}$$

Let

$$\begin{gathered} p(y=1|x) - p(y=0|x) \geqslant 0 \\ \text{namely} \\ \frac{p(x|y=1)-p(x|y=0)}{p(x|y=0)+p(x|y=1)} \geqslant 0 \end{gathered} \tag{3}$$

As

$$p(x|y=0)+p(x|y=1) \geqslant 0$$

Then the expression becomes as

$$\begin{aligned} p(x|y=1)-p(x|y=0) \geqslant 0 \\ \lambda_1 e^{-\lambda_1 x} - \lambda_0 e^{-\lambda_0 x} \geqslant 0 \end{aligned} \tag{4}$$

As $\lambda_i > 0$ and $\lambda_0 \neq \lambda_1$

$$(\lambda_0 - \lambda_1)x \geqslant ln(\frac{\lambda_1}{\lambda_0}) \tag{5}$$

x will be classified as class 1 when

$$\begin{cases} x \geqslant \frac{ln\lambda_1 - ln\lambda_0}{\lambda_0-\lambda_1}, & \text{if } \lambda_0 > \lambda_1 \\ x \leqslant \frac{ln\lambda_1 - ln\lambda_0}{\lambda_0-\lambda_1}, & \text{if } \lambda_0 < \lambda_1 \end{cases}$$

Q2:

As we discussed in the class, in the extreme situation, if the dataset is linear separable, the sigmoid function will tend to have the shape of step function. This means $\boldsymbol{w}->\infty$, as

$$\lim_{\boldsymbol{w}->\infty} \sigma(\boldsymbol{w}^T\boldsymbol{x}) = \lim_{\boldsymbol{w}->\infty} \frac{1}{1+e^{-\boldsymbol{w}^T\boldsymbol{x}}} = 1 \tag{6}$$

Similar to the linear regression problem, we can use a regularization term to penalize large weights.

Q3:

The sigmoid in 2-class has the form:

$$\begin{aligned} p(y=0|\boldsymbol{x}) =& \frac{1}{1+e^{-\boldsymbol{w}^T\boldsymbol{x}}} \\ p(y=1|\boldsymbol{x}) =& \frac{e^{-\boldsymbol{w}^T\boldsymbol{x}}}{1+e^{-\boldsymbol{w}^T\boldsymbol{x}}} \end{aligned} \tag{7}$$

The softmax function in 2-class:

$$\begin{aligned} p(y=c|\boldsymbol{x}) =& \frac{e^{\boldsymbol{w}_c^T\boldsymbol{x}}}{\sum_{c'} e^{\boldsymbol{w}_{c'}^T\boldsymbol{x}}} \\ p(y=0|\boldsymbol{x}) =& \frac{e^{\boldsymbol{w}_0^T\boldsymbol{x}}}{e^{\boldsymbol{w}_0^T\boldsymbol{x}} + e^{\boldsymbol{w}_1^T\boldsymbol{x}}} \\ =& \frac{1}{1+e^{(\boldsymbol{w}_1-\boldsymbol{w}_0)^T\boldsymbol{x}}} \\ p(y=1|\boldsymbol{x}) =& \frac{e^{\boldsymbol{w}_1^T\boldsymbol{x}}}{e^{\boldsymbol{w}_0^T\boldsymbol{x}} + e^{\boldsymbol{w}_1^T\boldsymbol{x}}} \\ =& \frac{e^{(\boldsymbol{w}_1-\boldsymbol{w}_0)^T\boldsymbol{x}}}{1+e^{(\boldsymbol{w}_1-\boldsymbol{w}_0)^T\boldsymbol{x}}} \end{aligned} \tag{8}$$

Assume $\boldsymbol{w}_1 - \boldsymbol{w}_0 = -\boldsymbol{w}$, then the expressions are equivalent.

Q4:

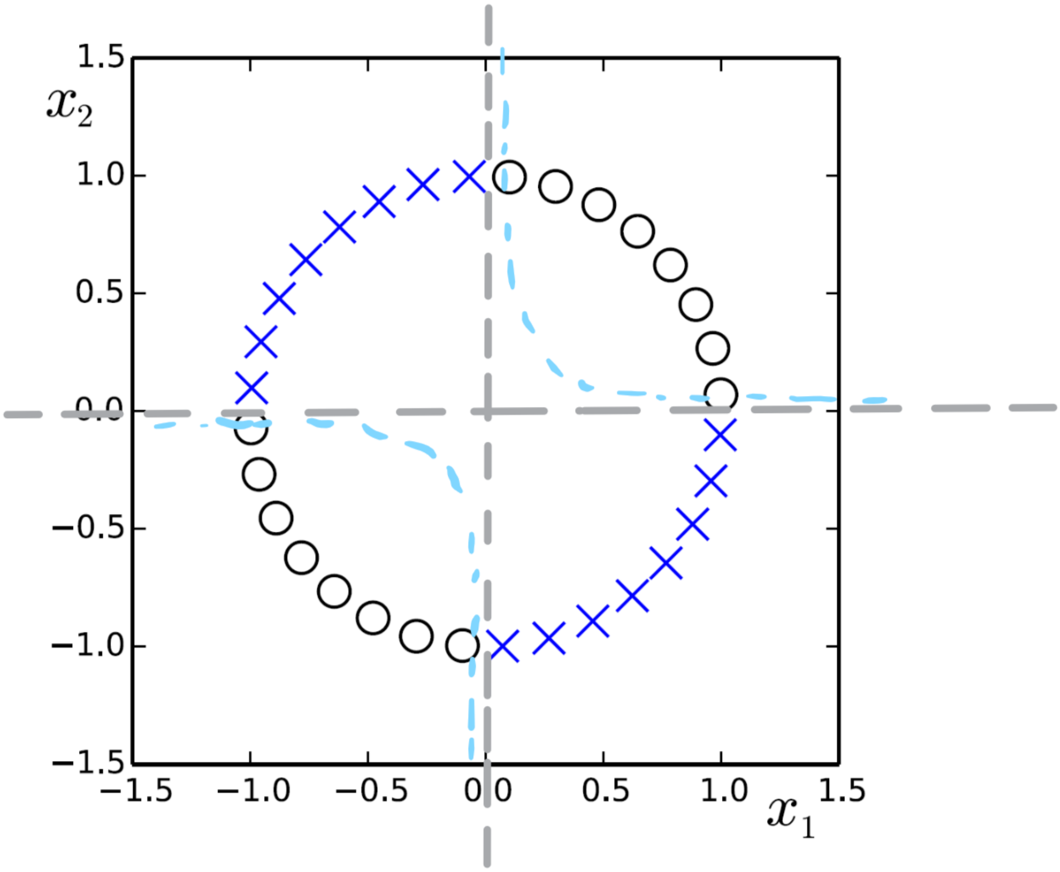

Figure 1: $\phi(x_1, x_2) = x_1 x_2$ separates the data

The basis equation can separate the data as in the figure.

$$\phi(x_1, x_2) = x_1 x_2$$