Problem 1. (Cross-ref: W7)

Recall the Naive Bayes model:

- For simplicity, we only consider binary features

$$x_i \in \{0,1\}, \quad \text{i.e.,} \quad \boldsymbol{x} \in \{0,1\}^d$$

- The generation model is

$$t \sim \text{Categorical}(\pi_1, \cdots, \pi_K)$$

$$\boldsymbol{x}_i | t = k \sim \text{Bernoulli}(p_{k,i})$$

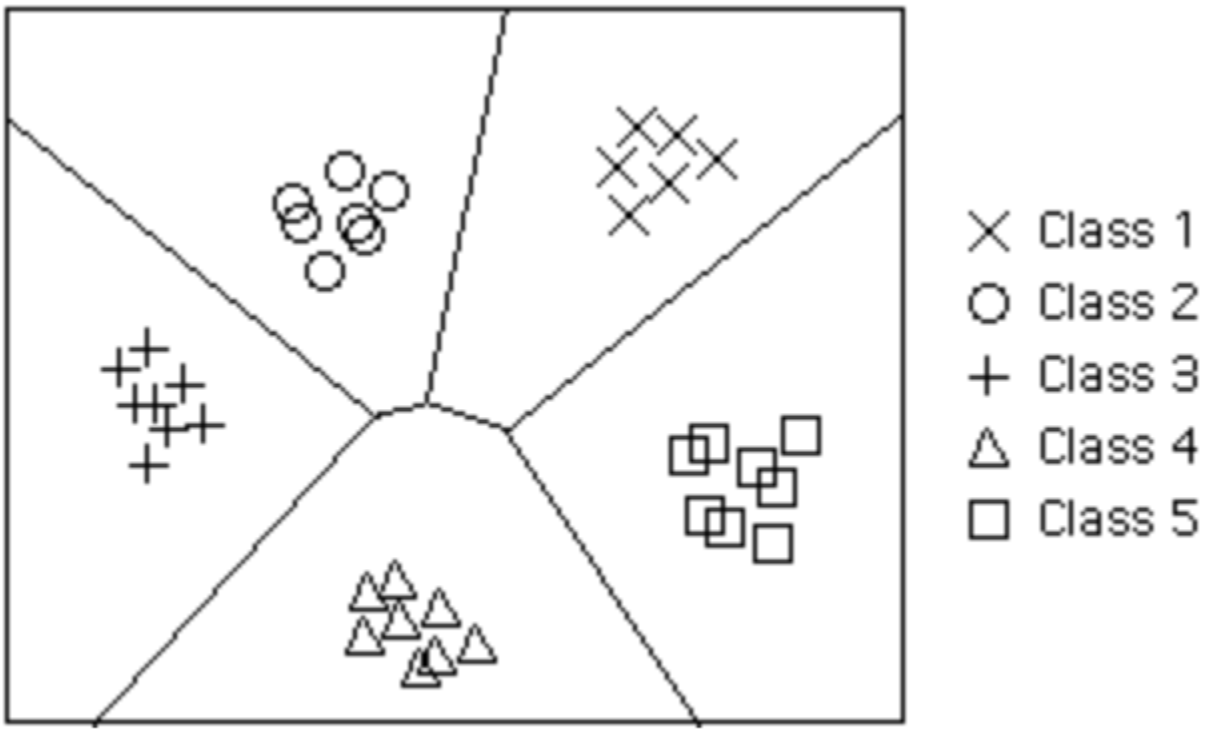
  Here: A Bernoulli distribution parametrized by π means that
  $\Pr[X = 1] = \pi$ and $\Pr[X = 0] = 1 - \pi$.
  It is a special case of categorical distributions in that only two cases are considered.

- Such a model can be used to represent a document in text classification. For example, the target indicates Spam or NotSpam. The feature indicates if a word in the vocabulary occurs in the document.

Show that the decision boundary is linear in $\boldsymbol{x}$.

Solution 1.

When we talk about decision boundaries, it is important to note that it is the pairwise decision boundary that we are talking about.



For example, the boundaries in the above figure are linear.

Recall the probability that a point belongs to class k:

$$\mathbb{P}(t=k \mid x_1, \cdots, x_d) = \frac{\mathbb{P}(x_1, \cdots, x_d \mid t=k)\mathbb{P}(t=k)}{\mathbb{P}(x_1, \cdots, x_d)} = \frac{\mathbb{P}(t=k)\Pi_{i=1}^{d}\mathbb{P}(x_i \mid t=k)}{\mathbb{P}(x_1, \cdots, x_d)}$$

If we had the probabilities, on the decision boundary between classes k and k' we would have

$$\mathbb{P}(t=k \mid x_1, \cdots, x_d) = \mathbb{P}(t=k' \mid x_1, \cdots, x_d)$$
$$\implies \log \mathbb{P}(t=k \mid x_1, \cdots, x_d) = \log \mathbb{P}(t=k' \mid x_1, \cdots, x_d)$$
$$\implies \log \mathbb{P}(t=k) + \sum_{i=1}^{d} \log \mathbb{P}(x_i \mid t=k) = \log \mathbb{P}(t=k') + \sum_{i=1}^{d} \log \mathbb{P}(x_i \mid t=k')$$

The true probabilities are unknown. Using the estimates instead, the equation for the boundary becomes

$$\log \hat{\pi}_k + \sum_{i=1}^{d} (x_i \log \hat{p}_{k,i} + (1-x_i)\log(1-\hat{p}_{k,i}))$$
$$= \log \hat{\pi}_{k'} + \sum_{i=1}^{d} (x_i \log \hat{p}_{k',i} + (1-x_i)\log(1-\hat{p}_{k',i}))$$

Rearranging the terms in the left-hand side and then defining $a_{k,0}$,..., $a_{k,d}$ gives

$$\log \hat{\pi}_k + \sum_{i=1}^{d} \Big(x_i(\log \hat{p}_{k,i} - \log(1-\hat{p}_{k,i})) + \log(1-\hat{p}_{k,i}) \Big)$$
$$= (\log \hat{\pi}_k + \sum_{i=1}^{d} \log(1-\hat{p}_{k,i})) + \sum_{i=1}^{d} x_i(\log \hat{p}_{k,i} - \log(1-\hat{p}_{k,i})) = a_{k,0} + \sum_{i=1}^{d} a_{k,i}x_i$$

We can do the same for the right-hand side. Then the decision boundary equation becomes

$$a_{k,0} + \sum_{i=1}^{d} a_{k,i}x_i = a_{k',0} + \sum_{i=1}^{d} a_{k',i}x_i \implies (a_{k,0} - a_{k',0}) + \sum_{i=1}^{d} (a_{k,i} - a_{k',i})x_i = 0$$

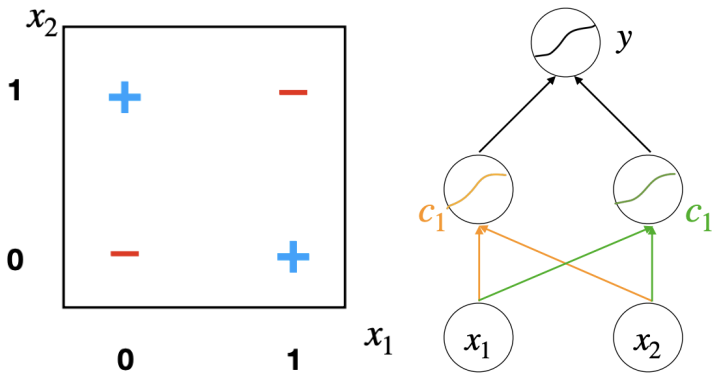
which is linear in x.

Problem 2.

Consider an XOR problem discussed in the lecture. It can be solved by stacking three logistic regression classifiers:

$$c_1 = \sigma(w_{11}x_1 + w_{12}x_2 + b_1)$$
$$c_2 = \sigma(w_{21}x_1 + w_{22}x_2 + b_2)$$
$$y = \sigma(w_1c_1 + w_2c_2 + b)$$

Give a set of weights (w_{11}, w_{12}, w_{21}, w_{22}, w_1, w_2, b_1, b_2, b) that solves the XOR problem.



Solution 2.

$w_{11} = 20, w_{12} = 20, w_{21} = 20, w_{22} = 20, w_1 = 20, w_2 = -20, b_1 = -10, b_2 = -30, b = -10$

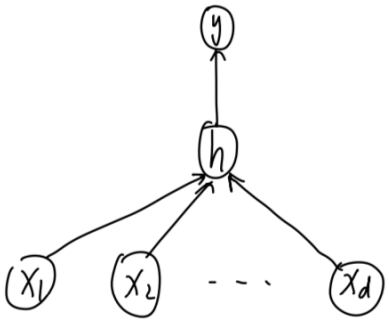
Note that if z is a large negative number then $\sigma(z)$ is approximately zero and if z is a large positive number then $\sigma(z)$ is approximately one. This way, c_1 becomes an OR gate, c_2 becomes an AND gate, and the second layer computes c_1 AND $\neg c_2$. The network as a whole becomes an XOR gate.

Problem 3.

Consider stacking two logistic regression units on top of the input $\boldsymbol{x} \in \mathbb{R}^d$ for classification. In other words, we have

$$h = \sigma(\boldsymbol{w}^\top \boldsymbol{x} + b)$$
$$y = \sigma(w_2 h + b_2)$$

where $\boldsymbol{w}, b, w_2, b_2$ are parameters. Show that such a classifier cannot achieve non-linearity.



Solution 3.

The decision boundary is:

$$y = 0.5 \implies \frac{1}{1+\exp(-(w_2 h + b_2))} = 0.5 \implies \exp(-(w_2 h + b_2)) = 1 \implies -(w_2 h + b_2) = 0 \implies h = -b_2/w_2$$

$$\frac{1}{1+\exp(-(\mathbf{w}\mathbf{x} + b))} = -\frac{b_2}{w_2} \implies 1 + \exp(-(\mathbf{w}\mathbf{x} + b)) = -\frac{w_2}{b_2} \implies -(\mathbf{w}\mathbf{x} + b) = \log(-1 - \frac{w_2}{b_2})$$

$$\implies \mathbf{w}\mathbf{x} + (b + \log(-1 - \frac{w_2}{b_2})) = 0$$

which is linear in x.

Problem 4.

Consider the following activation functions

$f_1(z) = z$

$f_2(z) = \sigma(z)$

$f_3(z) = \mathrm{ReLU}(z)$

where $\mathrm{ReLU}(z) = z$ if $z \geq 0$, or 0 otherwise.

Calculate the derivative of these functions wrt z.

1.

$$f_1'(z) = 1$$

2.

$$\begin{aligned}
f_2'(z) = \sigma(z)' &= \left(\frac{1}{1+\exp(-z)}\right)' \\
&= -\frac{1}{(1+\exp(-z))^2} \cdot (\exp(-z)) \cdot (-1) \qquad \text{(chain rule)} \\
&= \frac{\exp(-z)}{(1+\exp(-z))^2} \\
&= \frac{1+\exp(-z)-1}{(1+\exp(-z))^2} \\
&= \frac{1+\exp(-z)}{(1+\exp(-z))^2} - \frac{1}{(1+\exp(-z))^2} \\
&= \frac{1}{1+\exp(-z)} - \frac{1}{(1+\exp(-z))^2} \\
&= \frac{1}{1+\exp(-z)}\left(1 - \frac{1}{1+\exp(-z)}\right) \\
&= \sigma(z)(1-\sigma(z))
\end{aligned}$$

3.

$$f_3'(z) = \begin{cases} 1 & \text{if } z > 0 \\ 0 & \text{if } z < 0 \end{cases}$$

f_3(x) is not differentiable at z=0, because the left derivative is 0 and the right derivative is 1.

$$\lim_{\delta \to 0^+} \frac{f_3(\delta) - f_3(0)}{\delta} = \lim_{\delta \to 0^+} \frac{\delta - 0}{\delta} = 1$$

$$\lim_{\delta \to 0^-} \frac{f_3(\delta) - f_3(0)}{\delta} = \lim_{\delta \to 0^-} \frac{0 - 0}{\delta} = 0$$

But this is usually not a problem for gradient-based learning. Also, subgradients are sufficient to do whatever gradient is able to do. In this case, 0 or 1 can be the subgradient at 0.