THEODOR STOICAN
03725732

Assignment 5

Problem 1

a) $P(y|x)$ is a discrete distribution, since it has only 2 possible values:
- corresponding to $y=0$
- — " — $y=1$

In the most general case, $P(y|x)$ is a Bernoulli distribution, since we don't know if $P(y=1|x) = P(y=0|x)$.

b) x is classified as one iff $P(y=1|x) > \frac{1}{2}$

$$\Leftrightarrow \frac{P(x|y=1)\,P(y=1)}{P(x)} > \frac{1}{2}$$

$$\Leftrightarrow \frac{\lambda_1 e^{-\lambda_1 x} \cdot \cancel{\frac{1}{2}}}{P(x)} > \cancel{\frac{1}{2}}$$

$$\Leftrightarrow \frac{\lambda_1 e^{-\lambda_1 x}}{P(x, y=0) + P(x, y=1)} > 1$$

$$\Leftrightarrow \frac{\lambda_1 e^{-\lambda_1 x}}{P(x|y=0)P(y=0) + P(x|y=1)P(y=1)} > 1$$

$$\Leftrightarrow \frac{\lambda_1 e^{-\lambda_1 x}}{\lambda_0 e^{-\lambda_0 x} \cdot \frac{1}{2} + \lambda_1 e^{-\lambda_1 x} \cdot \frac{1}{2}} > 1$$

$$\Leftrightarrow \lambda_1 e^{-\lambda_1 x} > \frac{1}{2}\left(\lambda_0 e^{-\lambda_0 x} + \lambda_1 e^{-\lambda_1 x}\right)$$

$$\Leftrightarrow \cancel{\frac{1}{2}} \lambda_1 e^{-\lambda_1 x} > \cancel{\frac{1}{2}} \lambda_0 e^{-\lambda_0 x}$$

$$\Leftrightarrow \frac{e^{-\lambda_1 x}}{e^{-\lambda_0 x}} > \frac{\lambda_0}{\lambda_1}$$

$$\Leftrightarrow e^{x(\lambda_0 - \lambda_1)} > \frac{\lambda_0}{\lambda_1}$$

$$\Leftrightarrow x(\lambda_0 - \lambda_1) > \ln \lambda_0 - \ln \lambda_1$$

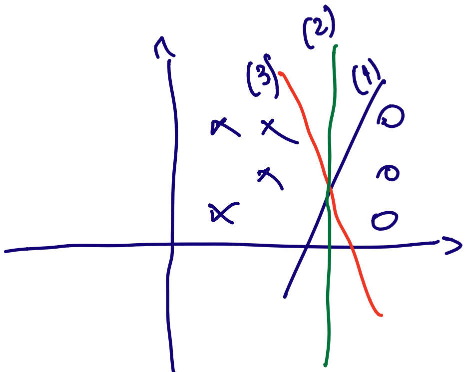
$$\Leftrightarrow \boxed{x > \frac{\ln \lambda_0 - \ln \lambda_1}{\lambda_0 - \lambda_1}}$$

Problem 2

Properties:

- the MLE estimate perfectly separates the dataset
- generally it is not unique

Example:



(1), (2), (3) separate the 2 classes perfectly. They are only 3 of the infinite number of lines that can do this. However, the greater w is, the greater the likelihood becomes too, since $\sigma(x) \to 1$ when $x \to \infty$. This is also the main problem with linearly separable datasets. One simple tweak would be to have a regularization term that penalizes large weights, such that we may stop when we have a hyperplane that separates the dataset perfectly and which has reasonably small weights.

Problem 3

For 2 classes, softmax would be:

$$\sigma(x)_1 = \frac{\exp(w_1 x + w_{10})}{\sum_{c=1}^{2} \exp(w_c x + w_{c0})}; \quad \sigma(x)_2 = \frac{\exp(w_2 x + w_{20})}{\sum_{c=1}^{2} \exp(w_c x + w_{c0})}$$

$$\sigma(x)_1 = \frac{1}{1+\frac{\exp(w_2 x + w_{20})}{\exp(w_1 x + w_{10})}} = \frac{1}{1+\exp\left(x(w_2 - w_1) + w_{20} - w_{10}\right)}$$

Similarly, $\sigma(x)_2 = \frac{1}{1+\exp\left(x(w_1 - w_2) + w_{10} - w_{20}\right)}$

We also know, as a property of softmax, that $\sigma(x)_1 = 1 - \sigma(x)_2$.

Now, let us consider $\sigma(x) = \frac{1}{1+\exp(wx + w_0)}$ the sigmoid function with
$w = w_2 - w_1$
$w_0 = w_{20} - w_{10}$

Whenever $\sigma(x) \geq 0.5 \Rightarrow$ class 1
$\sigma(x) < 0.5 \Rightarrow$ class 2

If $\sigma(x) \geq 0.5 \Rightarrow$ In softmax: $\sigma(x)_1 \geq 0.5$. Since $\sigma(x)_2 = 1 - \sigma(x)_1 \Rightarrow$
Softmax will classify the point as class 1.

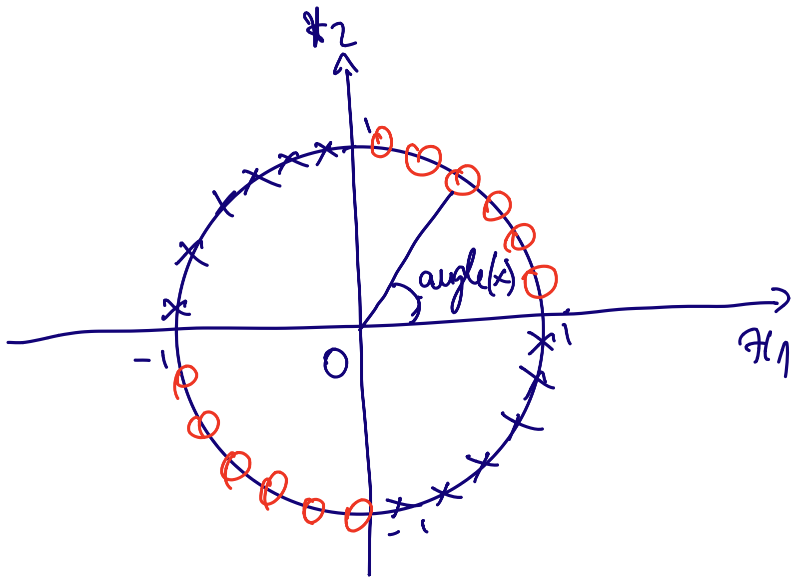
If $\sigma(x) < 0.5$, the logic follows the same pattern and softmax will yield class 2.

In a nutshell, we've found a sigmoid by combining the weights used by softmax and which performs the same classification as the latter.
Hence, softmax is equivalent to a sigmoid for the 2-class case.

Problem 4.

Intuition: The points from each of the 2 classes are in opposite sides of the circle. Hence, their "tan" value is either "≥ 0" or "< 0". So we could consider the coordinate of the new (1D) space as being the tan.

One such function would be $\phi(x) = (\tan(\text{angle}(x)))$



$\tan(\text{angle}(x)) \geq 0 \Rightarrow$ O

$\tan(\text{angle}(x)) < 0 \Rightarrow$ X