1 Linear classification

Problem 1: We want to create a generative binary classification model for classifying *nonnegative* one-dimensional data. This means, that the labels are binary ($y \in \{0, 1\}$) and the samples are $x \in [0, \infty)$.

We place a uniform prior on y

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

As our samples x are nonnegative, we use exponential distributions (and not Gaussians) as class conditionals:

$$p(x \mid y = 0) = \text{Expo}(x \mid \lambda_0) \qquad \text{and} \qquad p(x \mid y = 1) = \text{Expo}(x \mid \lambda_1),$$

where $\lambda_0 \neq \lambda_1$. Assume, that the parameters λ_0 and λ_1 are known and fixed.

a) What is the name of the posterior distribution $p(y \mid x)$? You only need to provide the name of the distribution (e.g., "normal", "gamma", etc.), not estimate its parameters.

Since $y \in \{0, 1\}$ the distribution is a Bernoulli distribution

b) What values of x are classified as class 1?
(As usual, we assume that the classification decision is $y_{predicted} = \arg\max_k p(y = k \mid x)$)

$\text{Expo}(x|\lambda) = \lambda \exp(-\lambda x)$

$p(y=1|x) = \frac{p(x|y=1)\,p(y=1)}{p(x)}$ Bayes' Rule

A data point x is classified as class 1 if

$\frac{p(y=1|x)}{p(y=0|x)} > 1$

$\frac{p(x|y=1)\,p(y=1)}{p(x)} > \frac{p(x|y=0)\,p(y=0)}{p(x)}$

$p(y=1) = p(y=0)$

$\frac{p(x|y=1)}{p(x|y=0)} > 1$ | log

$\log p(x|y=1) - \log p(x|y=0) > 0$

$$\log \lambda_1 - \lambda_1 x - \log \lambda_0 + \lambda_0 x > 0$$

$$\log \frac{\lambda_1}{\lambda_0} + (\lambda_0 - \lambda_1)x > 0$$

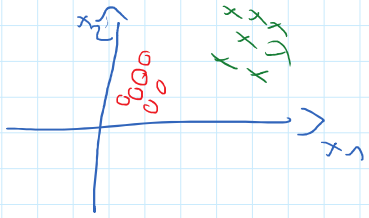
$$(\lambda_0 - \lambda_1)x > \log \frac{\lambda_0}{\lambda_1}$$

Caution: if $\lambda_0 - \lambda_1$ is negative, dividing by it will change the direction of the '>' relation.

x is classified as class 1 if $\begin{cases} x \in \left(\frac{\log \lambda_0 - \log \lambda_1}{\lambda_0 - \lambda_1}, \infty \right) & \text{if } \lambda_0 > \lambda_1 \\ x \in \left[0, \frac{\log \lambda_0 - \log \lambda_1}{\lambda_0 - \lambda_1} \right) & \text{if } \lambda_0 < \lambda_1 \end{cases}$

Problem 2: Assume you have a linearly separable data set. What properties does the maximum likelihood solution for the decision boundary $\boldsymbol{w}$ of ~~a linear~~ logistic regression model have? Assume that $\boldsymbol{w}$ includes the bias term.

What is the problem here and how do we prevent it?



From the lecture

$$p(y \mid w, X) = \prod_{i=1}^{N} p(y_i \mid x_i, w)$$

$$= \prod_{i=1}^{N} p(y=1 \mid x_i, w)^{y_i} \cdot \left(1 - p(y=1 \mid x_i, w)\right)^{1-y_i}$$

$$= \prod_{i=1}^{N} \sigma(w^T x_i)^{y_i} \left(1 - \sigma(w^T x_i)\right)^{1-y_i}$$

where $\sigma(a) = \frac{1}{1+\exp(-a)}$ is the sigmoid function

We define the error function to be the negative log-likelihood

$$E(w) = -\log p(y \mid w, X)$$

$$= -\sum_{i=1}^{N} y_i \log \sigma(w^T x_i) + (1-y_i) \log\left\{1 - \sigma(w^T x_i)\right\}$$

We are given that the dataset is linearly separable. Thus, we can find a w for which for all $1 \le i \le N$ it holds:

$$w^T x_i > 0 \quad \text{if } y_i = 1$$

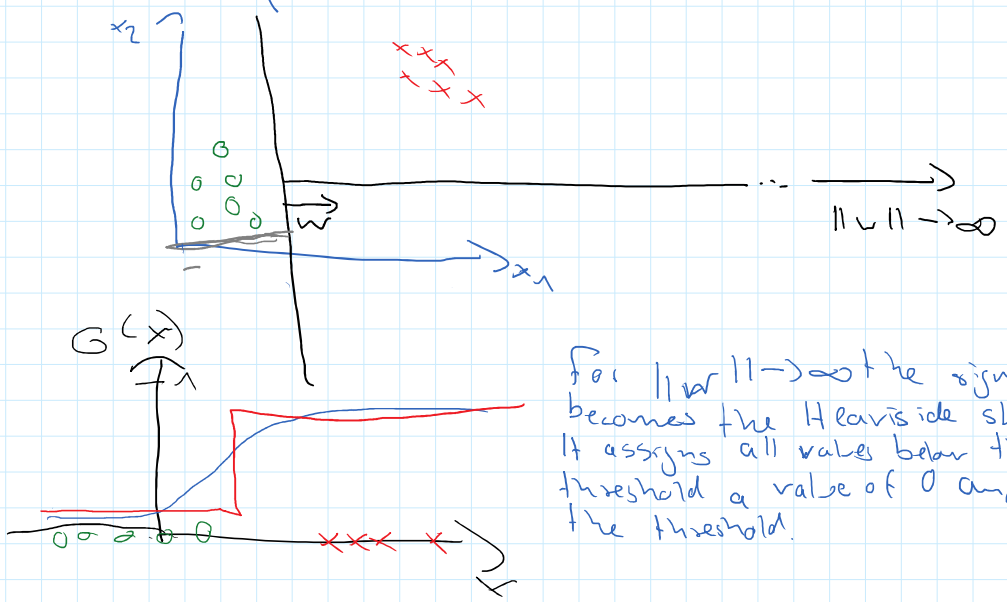
$$w^T x_i < 0 \quad \text{if } y_i = 0$$

$$E(w) = -\sum_{i=1}^{N} y_i \underbrace{\log \sigma(w^T x_i)}_{\max} + (1-y_i) \underbrace{\left(1 - \sigma(w^T x_i)\right)}_{\max}$$

$E(w)$: min

$\max + \max$

We achieve minimal error $E(w)$ for $\|w\| \to \infty$, i.e. the loss is not bounded



for $\|w\| \to \infty$ the sigmoid function becomes the Heaviside step function. It assigns all values below the decision threshold a value of 0 and 1 above the threshold.

To overcome this problem we can add a regularization term $\lambda \|w\|$ to the error function.

Problem 3: Show that the softmax function is equivalent to a sigmoid in the 2-class case.

$$p(y=1|x) = \frac{\exp(w_1^T x)}{\exp(w_1^T x) + \exp(w_0^T x)} = \frac{1}{1 + \frac{\exp(w_0^T x)}{\exp(w_1^T x)}}$$

$$= \frac{1}{1+\exp(w_0^T x - w_1^T x)} = \frac{1}{1+\exp(-(w_1 - w_0)^T x)}$$

$$= \frac{1}{1+\exp(-\hat{w}^T x)} = \sigma(\hat{w}^T x) \text{ where } \hat{w} = w_1 - w_0$$

Thus, the logistic regression model is not identifiable

We can add a constant $k \in \mathbb{R}$ to w_0 and w_1 and still have the same model

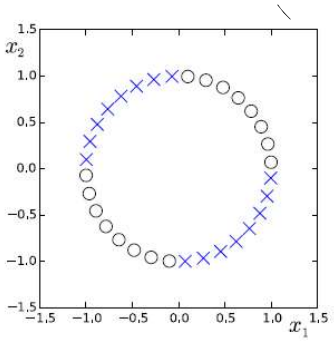
$$\hat{w} = (w_1 + k) - (w_0 + k) = w_1 - w_0 + k - k$$

The sigmoid function implicitly sets $w_0 = 0$ to overcome this issue.

HW Linear Classification (4)

Wednesday, 21 November 2018 14:11

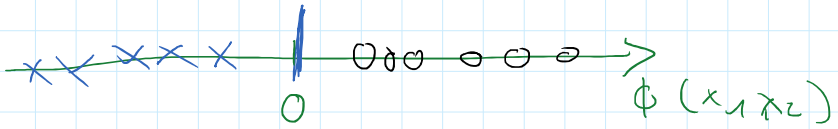
Problem 4: Which basis function $\phi(x_1, x_2)$ makes the data in the example below linearly separable (crosses in one class, circles in the other)?



x: one dimension is positive, the other one negative

o: both dimensions have the same sign

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



=> The dataset becomes linearly separable by applying the basis function $\phi(x_1, x_2) = x_1 x_2$.