MACHINE LEARNING HOMEWORK SHEET - 04

LINEAR CLASSIFICATION

NAME: MAHALAKSHMI SABANAYAGAM TUM ID: ge73yuw

DATE: 18.11.2018

Problem 1:

Generative binary classification for data $x \in [0,\infty)$ and labels $y \in \{0,1\}$

uniform prior $P(y=0) = P(y=1) = 1/2$

class conditionals $P(x|y=0) = Expo(x|\lambda_0) = \lambda_0 e^{-\lambda_0 x}$

$P(x|y=1) = Expo(x|\lambda_1) = \lambda_1 e^{-\lambda_1 x}$

a.) $P(y|x)$

$$P(y|x) \propto \underbrace{P(x|y)}_{\text{Likelihood}} \underbrace{P(y)}_{\text{uniform (so neglected)}}$$

$$\text{Likelihood } P(x|y) = \prod_{i=1}^{N} \underbrace{(\lambda_0 e^{-\lambda_0 x_i})^{1-y_i}}_{1 \text{ if } y_i = 1} \cdot \underbrace{(\lambda_1 e^{-\lambda_1 x_i})^{y_i}}_{1 \text{ if } y_i = 0}$$

$$P(y|x) \propto \prod_{i=1}^{N} \lambda_0^{1-y_i} \cdot \lambda_1^{y_i} \, e^{(-\lambda_0(1-y_i) - \lambda_1 y_i) x_i}$$

$$\propto k_1 e^{-k_2 x_i}$$

k_1, k_2 are constants.

$\therefore P(y|x)$ is exponentially distributed.

Lets find $P(y=1|x)$ and $P(y=0|x)$ for solving b.

$$P(y=k|x) = \frac{P(x|y=k) \cdot P(y=k)}{P(x)}$$

$$P(x) = \sum_{k=0,1} P(x|y=k) \cdot P(y=k)$$

$$P(y=1|x) = \frac{P(x|y=1)\,P(y=1)}{P(x|y=0)\,P(y=0) + P(x|y=1)\,P(y=1)}$$

$$= \frac{\lambda_1 e^{-\lambda_1 x}\,\frac{1}{2}}{\lambda_0 e^{-\lambda_0 x}\,\frac{1}{2} + \lambda_1 e^{-\lambda_1 x}\,\frac{1}{2}}$$

$$= \frac{1}{1 + \frac{\lambda_0}{\lambda_1} e^{(-\lambda_0+\lambda_1)x}}$$

$$= \frac{1}{1 + e^{-\left[\ln \frac{\lambda_1}{\lambda_0} - (\lambda_1-\lambda_0)x\right]}} \Rightarrow \text{Sigmoid function}$$

$$P(y=1|x) = \sigma\left(\ln \frac{\lambda_1}{\lambda_0} - (\lambda_1-\lambda_0)x\right)$$

$$P(y=0|x) = \frac{\lambda_0 e^{-\lambda_0 x}\,\frac{1}{2}}{\lambda_0 e^{-\lambda_0 x}\,\frac{1}{2} + \lambda_1 e^{-\lambda_1 x}\,\frac{1}{2}}$$

$$= \frac{1}{1 + \frac{\lambda_1}{\lambda_0} e^{(-\lambda_1+\lambda_0)x}}$$

$$= \frac{1}{1 + e^{-\left[\ln \frac{\lambda_0}{\lambda_1} - (\lambda_0-\lambda_1)x\right]}}$$

$$P(y=0|x) = \sigma\left(\ln \frac{\lambda_0}{\lambda_1} - (\lambda_0-\lambda_1)x\right).$$

b.) values of x classified as 1.

(ie) $P(y=1|x) > P(y=0|x)$

$$\frac{\lambda_1 e^{-\lambda_1 x} \cdot \cancel{1/2}}{\cancel{p(x)}} > \frac{\lambda_0 e^{-\lambda_0 x} \cdot \cancel{1/2}}{\cancel{p(x)}} \qquad p(x) \neq 0$$

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

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

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

For $x > \dfrac{\ln \frac{\lambda_0}{\lambda_1}}{\lambda_0 - \lambda_1}$, the target is 1.

Problem 3:

softmax function is sigmoid for 2 class case.

softmax function =

$$\sigma(x)_i = \frac{\exp(x_i)}{\sum_{k=1}^{k} \exp(x_i)}$$

for $k = 2$, & $i = 1$

$$\sigma(x)_1 = \frac{\exp(x_1)}{\exp(x_1) + \exp(x_2)}$$

$$= \frac{1}{1 + \frac{\exp(x_2)}{\exp(x_1)}}$$

Sigmoid fn:

$$\sigma(x) = \frac{\exp(x)}{1 + \exp(x)}$$

$$= \frac{1}{1 + e^{-x}}$$

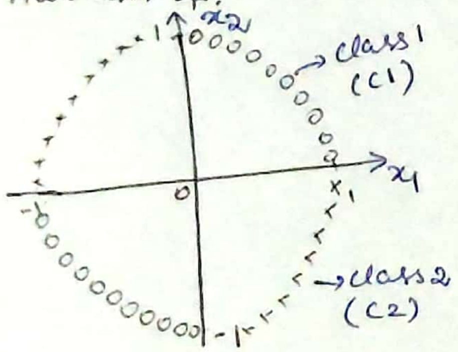
$$= \frac{1}{1 + e^{x_2 - x_1}}$$

Let $x_1 - x_2$ be x,

$$= \frac{1}{1 + e^{-x}} = \sigma(x)$$

$\therefore$ For 2 class case, Softmax function is equivalent to Sigmoid function.

Problem 4:



x_1	x_2	Class Label	
+	+	C1	← I Quadrant
−	−	C1	← III Quandrant
+	−	C2	← IV Quadrant
−	+	C2	← II Quadrant

Consider basis to be $(x_1, x_2, x_1 x_2)$

x_1	x_2	$x_1 x_2$	Label	
+	+	+	C1	All +ve values of $x_1 x_2$ belong to C1
−	−	+	C1	
+	−	−	C2	And −ve values of $x_1 x_2$ belong to C2
−	+	−	C2	

$\therefore$ When we take the given data (x_1, x_2) to a higher dimension (3D) with $(x_1, x_2, x_1 x_2)$ as basis it becomes linearly separable.

The separating hyperplane will be $x_1 x_2 = 0$

The positive side of the plane (ie) $x_1 x_2 > 0$ is class 1 & negative side $x_1 x_2 < 0$ is class 2

Problem 2:

Linearly separable data.

Maximum Likelihood solution for the decision boundary w of logistic regression model properties:

Likelihood of logistic regression:

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

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

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

Negative log likelihood,

$$= -\sum_{i=1}^{N} y_i \ln \sigma(w^T x_i) + (1-y_i)\ln\left(1-\sigma(w^T x_i)\right) \quad ; \; \sigma(a) = \frac{1}{1+e^{-a}}$$

$$\ln(\sigma(a)) = -\ln(1+e^{-a})$$

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

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

$$= -\sum_{i=1}^{N} \cancel{-y_i \ln(1+e^{-w^T x_i})} \mp w^T x_i + y_i w^T x_i - \ln(1+e^{-w^T x_i}) \cancel{+ y_i \ln(1+e^{-w^T x_i})}$$

$$= -\sum_{i=1}^{N} -w^T x_i + y_i w^T x_i - \ln(1+e^{-w^T x_i})$$

Finding gradient w.r.t. w,

$$\nabla_{(w)} = -\sum_{i=1}^{N} -x_i^T + y_i x_i^T + \frac{e^{-w^T x_i}}{1+e^{-w^T x_i}} x_i^T$$

$$= -\sum_{i=1}^{N} \left[\frac{-1 \cancel{- e^{-w^T x_i}} \cancel{+ e^{-w^T x_i}}}{1+e^{-w^T x_i}}\right] x_i^T + y_i x_i^T$$

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

The maximum likelihood solution for w is very much similar to gradient of error function for the linear regression model.

It severely over fits the data.

This can be prevented by the right choice of optimization algorithm and parameter initialization.