4.19) Generative classifier

$$p_\theta(y=c/x) = \frac{P_\theta(y=c)\, P_\theta(x/y=c)}{\sum_{c'} P_\theta(y=c')\, P_\theta(x/y=c')}$$

In Gaussian discriminant, we model

$$P(y) = \mathrm{Mul}(\pi) \quad P(x/y=c) = N(\mu_c, \Sigma_c)$$

$$\theta = [\pi_{1:c}, \mu_{1:c}, \Sigma_{1:c}]$$

In given problem, we have only two classes $c=2$ and $\Sigma_1 = K\Sigma_0$, $K>0$, then we can write

$$P(y=1/x) = \frac{\pi_1 \cdot |2\pi\Sigma_1|^{-1/2} e^{-\frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1)}}{\sum_{i=0}^{1} \pi_i \cdot |2\pi\Sigma_i|^{-1/2} e^{-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)}}$$

We'll simplify numerator & denominator seperatly

by substituting ~~K~~ $\Sigma_1 = K\Sigma_0$, $K>0$

$$\text{Numerator of } p(y=1/x) = \pi_1 \cdot |2\pi \cdot (K\Sigma_0)|^{-1/2} \cdot e^{-\frac{1}{2}(x-\mu_1)(K\Sigma_0)^{-1}(x-\mu_1)}$$

$$= \frac{1}{\sqrt{K}} \cdot \pi_1 \cdot |2\pi\Sigma_0|^{-1/2} \cdot e^{-\frac{1}{2K}(x-\mu_1)\Sigma_0^{-1}(x-\mu_1)}$$

Denominator of $p(y=1/x) =$

$$= \pi_1 \cdot |2\pi(K\Sigma_0)|^{-1/2} \cdot e^{-\frac{1}{2}(x-\mu_1)(K\Sigma_0)^{-1}(x-\mu_1)} \cancel{+\pi_0 |2\pi\Sigma_0|^{-1/2}}$$

$$+ \pi_0 \cdot |2\pi\Sigma_0|^{-1/2} \cdot e^{-\frac{1}{2}(x-\mu_0)\Sigma_0^{-1}(x-\mu_0)}$$

Scanned by CamScanner

$$= \cancel{\frac{1}{\sqrt{K}}} (\pi_1 \cdot \cancel{|2\pi\Sigma_0|^{-1}}$$

$$= |2\pi\Sigma_0|^{-1/2} \cdot e^{-\frac{1}{2}x^T\Sigma_0 x} \left[ \frac{\pi_1}{\sqrt{K}} \cdot e^{\frac{1}{K}(\mu_1^T\Sigma_0^{-1}x - \frac{1}{2}\mu_1^T\Sigma_0\mu_1)} \right.$$

$$\left. + \pi_0 \cdot e^{(\mu_0^T\Sigma_0^{-1}x - \frac{1}{2}\mu_0^T\Sigma_0\mu_0)} \right]$$

After we cancel first term i.e $|2\pi\Sigma_0|^{-1/2} \cdot e^{-\frac{1}{2}x^T\Sigma_0 x}$

both in numerator & denominator we get.

$$= \frac{\left[ \frac{\pi_1}{\sqrt{K}} \cdot e^{1/K(\mu_1^T\Sigma_0^{-1}x - \frac{1}{2}\mu_1^T\Sigma_0\mu_1)} \right]}{\frac{\pi_1}{\sqrt{K}} \cdot e^{1/K(\mu_1^T\Sigma_0^{-1}x - \frac{1}{2}\mu_1^T\Sigma_0\mu_1)} + \pi_0 \cdot e^{(\mu_0^T\Sigma_0^{-1}x - \frac{1}{2}\mu_0^T\Sigma_0\mu_0)}}$$

lets simplify this to by substituting

$$\beta_1 = \Sigma^{-1}\mu_1 \qquad \beta_0 = \Sigma^{-1}\mu_0$$

$$\gamma_1 = -\frac{1}{2}\mu_1^T\Sigma_0\mu_1 + \log\pi_1 \qquad \gamma_0 = -\frac{1}{2}\mu_0^T\Sigma_0\mu_0 + \log\pi_1$$

$$= \frac{e^{1/K(\beta_1^T x + \gamma_1 - \frac{\log K}{2})}}{e^{1/K(\beta_1^T x + \gamma_1 - \frac{1}{2}\log K)} + e^{(\beta_0^T x + \gamma_0)}}$$

$$= \frac{1}{1 + e^{\beta_0^T x + \gamma_0 - 1/K(\beta_1^T x + \gamma_1 + \frac{1}{2}\log K)}}$$

$$= \frac{1}{1+e^{(\beta_0 - 1/k\,\beta_1)^T x + (\gamma_0 - 1/k\,\gamma_1) + \frac{\log k}{2k}}}$$

$$= \text{sigm}\left( (\beta_{1/k} - \beta_0)^T x + (1/k\,\gamma_1 - \gamma_0) + \frac{\log k}{2k} \right).$$

So, Even $\Sigma_1 = k\Sigma_0$, the discriminative ~~surface~~ function is still sigmoid function. and boundary is still linear.

4.21) Gaussian decision boundaries:

$$P(x|y=j) = N(x|\mu_j, \sigma_j) \quad j=1,2$$

So, we have two univariate distributions with parameters

$$\mu_1 = 0 \qquad \mu_2 = 1$$
$$\sigma_1^2 = 1 \qquad \sigma_2^2 = 10^6.$$

a) Decision region is given by

$$R_1 = \{x : P(x|\mu_1, \sigma_1) \geq P(x|\mu_2, \sigma_2)\}$$

To get solutions for this inequality, let use distribution formula and substitute the parameters.

$$\frac{1}{\sqrt{2\pi}\,\sigma_1} \cdot e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} \geq \frac{1}{\sqrt{2\pi}\,\sigma_2} \cdot e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}$$

$$\frac{1}{1} \cdot e^{-\frac{(x-0)^2}{2\cdot 1}} \geq \frac{1}{10^3} \cdot e^{-\frac{(x-1)^2}{2\cdot 10^6}}$$

$$e^{-\frac{x^2}{2} + 3\ln 10} \geq e^{-\frac{(x-1)^2}{2\cdot 10^6}}$$

take log on both sides

$$-\frac{x^2}{2} + 3\ln 10 \geq -\frac{(x^2 - 2x + 1)}{2\cdot 10^6}$$

$$-10^6 \cdot x^2 + 6\cdot 10^6 \ln 10 \geq -x^2 + 2x - 1$$

$$(10^6 - 1)x^2 + 2x - 1 - 6\cdot 10^6 \ln 10 \leq 0.$$

This is quadratic equation with solution

$$x \cdot -\frac{-2 \pm \sqrt{4 - 4\cdot(10^6-1)\cdot(-1 - 6\cdot 10^6 \ln 10)}}{2\cdot(10^6-1)} \leq 0$$

$$x \cdot -\frac{-1 \pm \sqrt{1 + (10^6-1)(1 + 6\cdot 10^6 \ln 10)}}{(10^6 - 1)} \leq 0$$

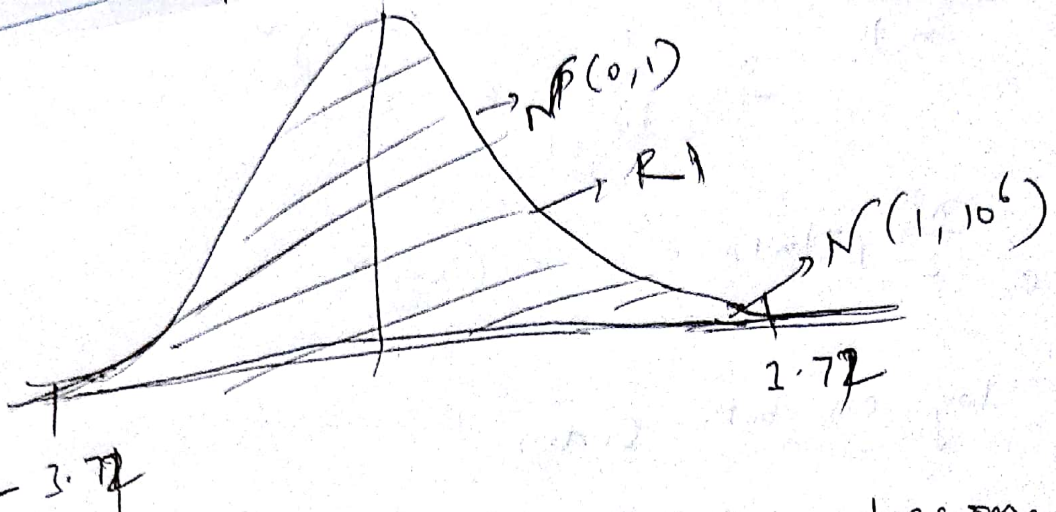
$$x \in [-3.72, 3.72]$$

So, Region where $\{x : p(x|\mu_1, \sigma_1) \geq p(x|\mu_2, \sigma_2)\}$

in $x \in [-3.72, 3.72]$

To visualize, we can plot both distributions and

see where they intersect each other.




b). if $\sigma_2 = 1$, The equation becomes

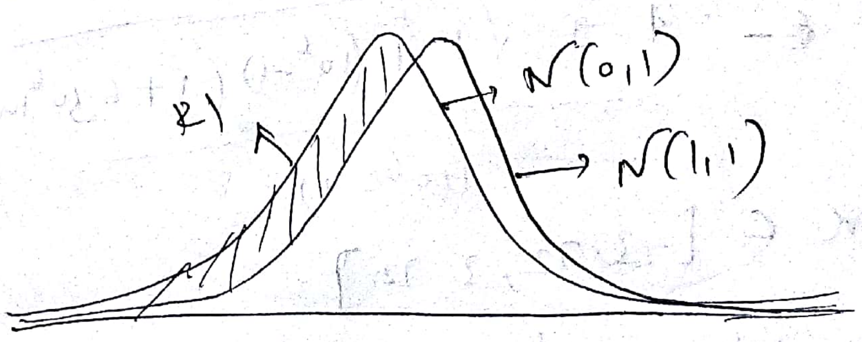
$$e^{-\frac{x^2}{2}} \geq e^{-\frac{(x-1)^2}{2}}$$

$$\frac{(x-1)^2}{2} - \frac{x^2}{2} \geq 0$$

$$-2x+1 \geq 0$$

$$x \leq 1/2$$

Then $R1 = \{x \in R;\ x \leq 1/2\}$




4.22) It's solved as programming exercise.

Please refer Exercise-4.22.ipynb file.

Result: Class for X1 : 1

Class for X2 : 2

4.20) a) Lets compare GaussI and Linlog.

In GaussI, we are modeling class conditional densities with ~~uniform prior~~ ~~and~~ identity matrix covariance $P(x/y=c) = N(x/\mu_c, I)$ and prior as uniform. So, the derivation leads to

$$p(y=c/x) = \frac{e^{\beta_c^T x + r_c}}{\sum_{c'} e^{\beta_c^T x + r_c}}$$

Here $\beta_c = \Sigma^{-1}\mu_c = \mu_c \quad (\because \Sigma_c = I)$

and $r_c = -\frac{1}{2}\mu_c^T \Sigma^{-1}\mu_c = -\frac{1}{2}\|\mu_c\|^2$

So $$p(y=c/x) = \frac{e^{\mu_c^T x - \frac{\|\mu_c\|^2}{2}}}{\sum_{c'} e^{\mu_c^T x - \frac{\|\mu_c\|^2}{2}}} \rightarrow (M)$$

In ~~linear~~ Linlog (linear logistic regression) we model conditional likelihood as

$$p(y=c/x) = \frac{e^{w_c^T x}}{\sum_{c'} e^{w_c^T x}} \rightarrow (M')$$

This equation ~~eaea~~ is exactly like above GaussI with constant parameter $\left(w_0 = \frac{-\|\mu_c\|^2}{2}\right)$.

So, the conditional log-likelihood for M and M' will generate same performance (equal).

$$\boxed{L(M) = L(M')}$$

b) lets compare GaussX, QuadLog.

In GaussX, we are modeling class conditional densities $p(x/y=c) = N(x/\mu_c, \Sigma_c)$ and prior as uniform

so, posterior equation becomes

$$p(y=c/x) = \frac{e^{-\frac{1}{2}x\Sigma_c^{-1}x + \mu_c^T\Sigma_c^{-1}x - \frac{1}{2}\mu_c^T\Sigma_c^{-1}\mu_c}}{\Sigma_{c'} e^{-\frac{1}{2}x\Sigma_c^{-1}x + \mu_c^T\Sigma_c^{-1}x - \frac{1}{2}\mu_c^T\Sigma_c^{-1}\mu_c}}$$

This is nothing but sigmoid function in quadratic form

$$p(y=c/x) \propto e^{w_c^T\phi(x)} \quad - \; M$$

where $\phi(x)$ is quadratic equation of $x$.

In ~~lo~~ QuadLog (logistic regression with quadratic features) we model

$$p(y=c/x) \propto e^{w_c^T\phi(x)} \quad - \; M'$$

with $\phi(x)$ as quadratic function of features $x$.

So, log-liklihood for GaussX and QuadLog $L(M)$ and $L(M')$ will be same.

Scanned by CamScanner

c) compare linlog, Quadlog.

In linlog (logistic regression with linear features)

we model posterior as

$$P(y=c/x) \propto e^{w_c^T x} \quad -\text{(M)}$$

In Quadlog (logistic regression with quadratic features)

we model posterior as

$$P(y=c/x) \propto e^{w_c^T \phi(x)} \quad -\text{(M')}$$

$\phi(x)$ - quadratic function of feature $x$.

If we compare $L(M)$ and $L(M')$ (log likelihood of both)

we can say $L(M')$ does better in all datasets where $L(M)$ does good. But ~~L(M)~~ $M'$ will fit better for non-linear data but not $M$. So, we can say

$$L(M') \geq L(M)$$

d) compare GaussI, Quadlog.

As we derived before GaussI models data directly with $(\mu_c, -\frac{\|\mu_c\|^2}{2})$ as weights, so its same as linlog. But as discussed above Quadlog models features as quadratic function, so it fits ~~bad~~ both non-linear and linear data (when coefficient of $x^2=0$)

So $L(M)$ $L(M') \geq L(M)$

8.6) $l_2$ regularized logistic regression

$$J(w) = -\ell(w, D_{train}) + \lambda \|w\|^2$$

$$\ell(w, D) = \frac{1}{|D|} \sum_{i \in D} \log \sigma(\underbrace{y_i x_i^T w}_{= \theta_i})$$

a) The gradient of $J(w)$

$$g = \frac{dJ}{dw} = \sum_i (\theta_i - y_i) x_i = X^T(\theta - y)$$

Hessian for $J(w)$

$$H = \frac{d}{dw} g(w)^T = \sum_i (\nabla_w \theta_i) x_i^T$$

$$= \sum_i \theta_i (1-\theta_i) x_i x_i^T$$

$$= X^T S X.$$

As $H$ is Heissian, we'll have unique minim

for $J(w)$

So, $J(w)$ has multiple solutions? FALSE.

b) $\hat{w} = \arg\min_w J(w)$ is global optimum.

As we are deliving $\hat{w}$ by gradient descent, we get aproximate solution not exact one

So, $\hat{w}$ will not be sparse as its not generated analytically.

Scanned by CamScanner

b) So, $\hat{w}$ is sparse ? FALSE

c) If training data is linearly seperable then $\|w\| \to \infty$ (without regularization term)

So, if $\lambda = 0$ then some weights may become infinite ? TRUE

d) $l(\hat{w}, D_{train})$ ~~de~~ always increases as we increase $\lambda$ ?

FALSE.

When we increase $\lambda$, then we are penalizing 'w' to fit to train data, so log-liklihood for $D_{train}$ will decrease.

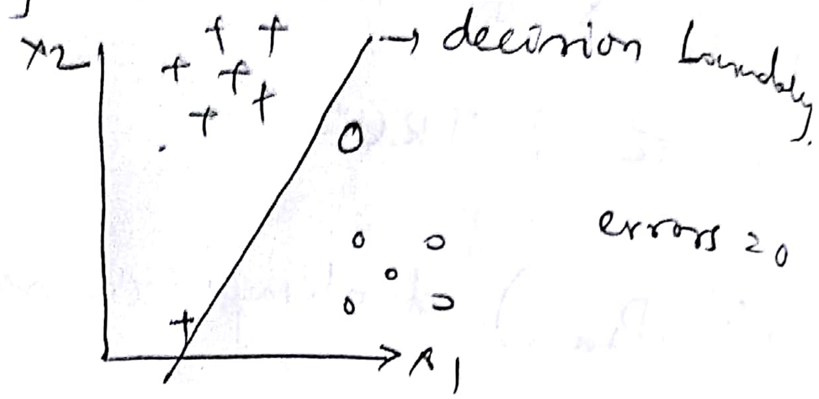
e) $l(\hat{w}, D_{test})$ always increases as we increase $\lambda$ ?

~~TRUE~~ FALSE

When we increase $\lambda$, then we are making model smoother to fit to test data. But when we increase further, then it underfits the data and model becomes too simple so $l(\hat{w}, D_{test})$ will start decreasing.

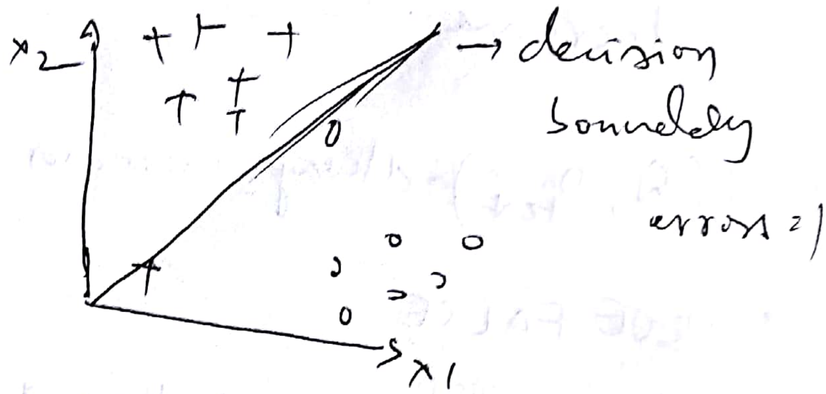
8.1) a) $J(w) = -l(w, D_{train})$

As we are not regularizing the weights, w values will come such that they explain the data perfectly



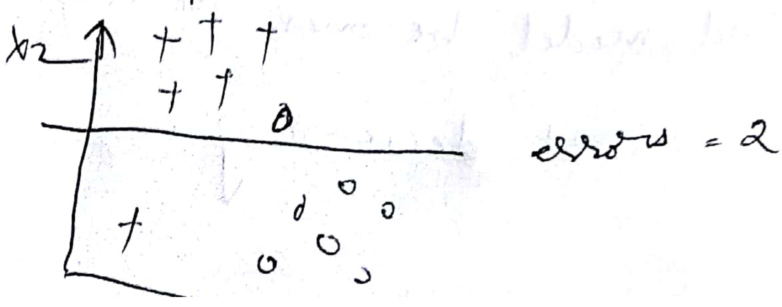

b) $J_0(w) = -l(w, D_{train}) + \lambda w_0^2$

Here we are regularizing only $w_0$. So the offset of decision boundary becomes 0.

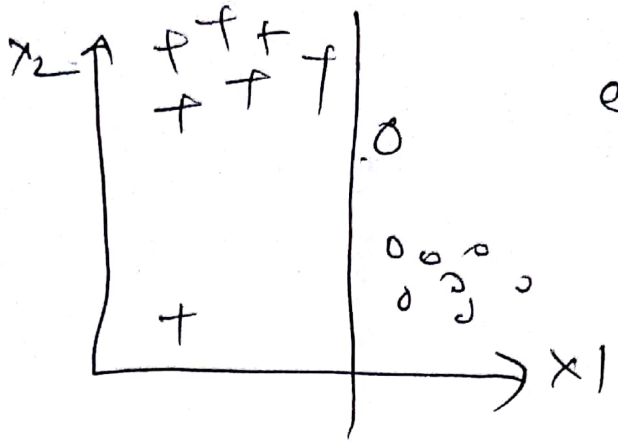



c) $J_1(w) = -l(w, D_{train}) + \lambda w_1^2$

Here, the decision boundary will be parallel to x1, as $w_1 = 0$



d) Similarly when we try to regularize only $w_2$ then we get




errors = 0

Scanned by CamScanner