Look at one class at a time instead of 2

Discriminative Learn $p(y|x)$ (or learn $h_{\theta}(x) = \{0,1\}$)

Generative build model of each first Learn $p(x|y)$ & $p(y)$

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

$$p(x) = p(x|y=1)p(y=1) + p(x|y=0)p(y=0)$$

GDA Gaussian Discriminant Analysis

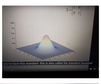
Suppose x continuous $x \in R^n$ (instead of $n+1$)

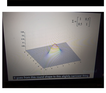
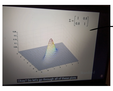
Assume $p(x|y)$ is Gaussian

Multi-variate Gaussian $z \sim N(\vec{\mu}, \Sigma)$ ($\vec{\mu} \in R^n$, Σ covariance $R^{n \times n}$)

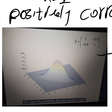
$$E[z] = \mu, \quad Cov[z] = E[(z-\vec{\mu})(z-\vec{\mu})^T]$$
$$= E[zz^T] - (Ez)(Ez)^T$$

$$p(z) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\vec{\mu})^T\Sigma^{-1}(x-\vec{\mu})}$$



→ 2, & 2 positively correlated



→ negatively correlated

GDA

$$p(x|y=0) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)}$$

$p(x|y=1) =$ ..... same Σ but different μ

$p(y) = \phi^y(1-\phi)^{1-y}$

Training set $\{(x^{(i)}, y^{(i)})\}_{i=1}^m$

Goal max Joint likelihood

Discriminative $\ell(\theta) = \prod_{i=1}^m p(y^{(i)}|x^{(i)}, \theta)$

$$\ell(\phi, \mu_0, \mu_1, \Sigma) = \prod_{i=1}^m p(x^{(i)}, y^{(i)}; \phi, \mu_0, \mu_1, \Sigma)$$

$$\phi = \frac{\sum_{i=1}^m y^{(i)}}{m} = \frac{\sum_{i=1}^m 1\{y^{(i)}=1\}}{m}$$

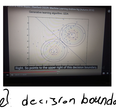
$$\mu_0 = \frac{\sum_{i=1}^m 1\{y^{(i)}=1\}x^{(i)}}{\sum 1\{y^{(i)}=0\}}$$ ← sum of feature vectors for $y=0$

$$\mu_1 = \frac{\sum 1\{y^{(i)}=1\}x^{(i)}}{\sum 1\{y^{(i)}=1\}}$$ ← # with $y=0$

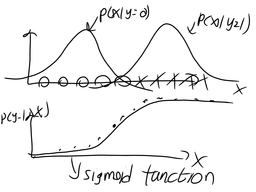
$$\Sigma = \frac{1}{m}\sum_{i=1}^m (x^{(i)}-\mu_{y^{(i)}})(x^{(i)}-\mu_{y^{(i)}})^T$$

prediction

$$\text{return } \arg\max_y p(y|x) = \arg\max \frac{p(x|y)p(y)}{p(x)} = \arg\max p(x|y)p(y)$$



Σ same decision boundary linear ? end up with quadratic terms in logistic function



Both discriminative / GDA use sigmoid function

GDA assess
$x|y=0 \sim N(\mu_0, \Sigma)$
$x|y=1 \sim N(\mu_1, \Sigma)$
$y \sim Ber(\phi)$
$\Rightarrow$ Logistic regression $p(y=1|x) = \frac{1}{1+e^{-\theta x}}$

Thus GDA stronger assumptions

If $p(x|y)$ gaussian → GDA better

Generally logistic more robust

eg $x|y=1 \sim Poisson(\lambda_1)$, $x|y=0 \sim Poisson(\lambda_0)$ $\Rightarrow p(y=1|x)$ is logistic

If they both same exp family, then logistic

GDA → computation efficient

LR → generally better

Naive Bayes Algorithm

Feature vector $X = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$ a vocab, 10000 = n

$x_i = \text{Ind.cator}\{\text{word } i \text{ appears}\}$ 3 must

want to model $p(x|y)$, $p(y)$

2^{10000} of x

Assume x_i's are conditionally independent

$$p(x_1, x_{10000}|y) = p(x_1|y)\,p(x_2|x_1, y)\,p(x_3|x_1, x_2, y) \cdots p(x_{10000}|x_1, \ldots, y)$$
$$= p(x_1|y)p(x_2|y) \cdots p(x_{10000}|y)$$
$$= \prod_{i=1}^n p(x_i|y)$$

Parameters

$\phi_{j|y=1} = p(x_j=1|y=1)$

$\phi_{j|y=0} = p(x_j=1|y=0)$

$\phi_y = p(y=1)$

Joint likelihood

$\ell(\phi_y, \phi_{j|y})$

MLE $\phi_y = \frac{\sum_{i=1}^m 1\{y^{(i)}=1\}}{m}$

$\phi_{j|y=1} = \frac{\sum 1\{x_j^{(i)}=1, y^{(i)}=1\}}{\sum 1\{y^{(i)}=1\}}$