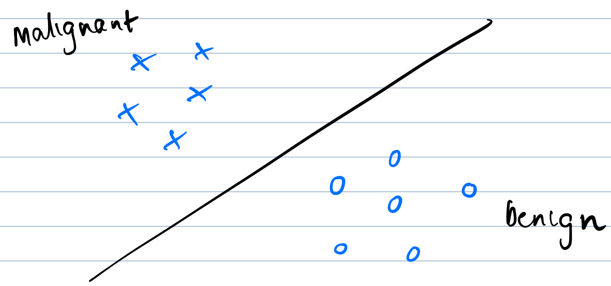
Announcements : Wed Oct 6

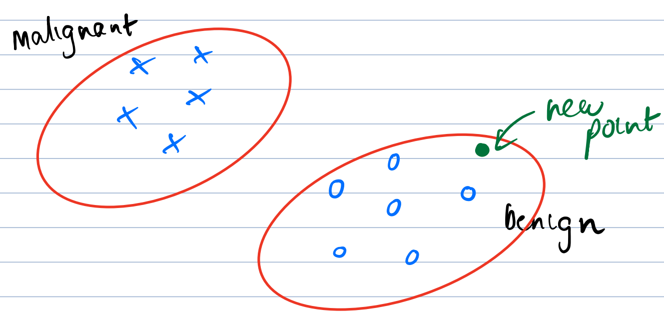
- PSET 1 due ~~Thu Oct 7~~, 11:59 pm
- working on matching project groups with TAs

Generative Learning Algorithms

- Gaussian Discriminant Analysis (GDA)
- Generative & Discriminative Comparison
- Naive Bayes



Discriminative Algorithm



Generative Algorithm

Discriminative Learning Algorithm

Learns $p(y|x)$

or learns $h_\theta(x) = \begin{cases} 0 \\ 1 \end{cases}$ directly

Generative Learning Algorithm

Learns $p(x|y)$

features → x, class → y

$p(y)$ ← class prior

Bayes Rule

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

$$p(x) = p(x|y=1)\, p(y=1) + p(x|y=0) \cdot p(y=0)$$

Gaussian Discriminant Analysis (GDA)

Suppose $x \in \mathbb{R}^d$ (drop $x_0 = 1$ convention)

Assume $p(x|y)$ is Gaussian

$$Z \sim N(\underbrace{\vec{\mu}}_{\mathbb{R}^d}, \underbrace{\Sigma}_{\mathbb{R}^{d\times d}}) \qquad Z \in \mathbb{R}^d \quad (z_1, z_2 \ldots z_d)$$

$$\mathbb{E}[Z] = \vec{\mu}$$

$$\mathrm{Cov}[Z] = \mathbb{E}[(Z-\mu)(Z-\mu)^T]$$

$$= \mathbb{E}[ZZ^T] - (\mathbb{E}[Z])(\mathbb{E}[Z])^T$$

$$p(z) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

GDA model

$$p(x \mid y=0) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_0)^T \Sigma_0^{-1} (x-\mu_0)\right)$$

$$p(x \mid y=1) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1)\right)$$

Parameters : $\mu_0, \mu_1, \Sigma, \phi$

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

$$P(y=1) = \phi$$

$$\underbrace{\mu_0 \quad \mu_1}_{\mathbb{R}^d} \qquad \underset{\mathbb{R}^{d \times d}}{\Sigma} \qquad \underset{\mathbb{R}\ \in [0,1]}{\phi}$$

Training Set $\{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$

Joint Likelihood

$$\mathcal{L}(\phi, \mu_0, \mu_1, \Sigma) = \prod_{i=1}^{n} P(x^{(i)}, y^{(i)}; \phi, \mu_0, \mu_1, \Sigma)$$

$$= \prod_{i=1}^{n} P(x^{(i)} \mid y^{(i)}) \cdot P(y^{(i)})$$

Discriminative:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} P(y^{(i)} \mid x^{(i)}; \theta)$$

Conditional Likelihood

Maximum Likelihood Estimation

$$\max_{\phi, \mu_0, \mu_1, \Sigma} \ \ell(\phi, \mu_0, \mu_1, \Sigma) = \log \mathcal{L}(\ldots)$$

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

$$\mathbb{1}\{\text{true}\} = 1$$

$$\mathbb{1}\{\text{false}\} = 0$$

$$\mu_0 = \frac{\sum_{i=1}^{n} \mathbb{1}\{y^{(i)}=0\} \cdot x^{(i)}}{\sum_{i=1}^{n} \mathbb{1}\{y^{(i)}=0\}}$$

← sum of feature vectors for examples w. $y^{(i)}=0$

← # examples w. $y^{(i)}=0$

$$\mu_1 = \frac{\sum_{i=1}^{n} \mathbb{1}\{y^{(i)}=1\} \cdot x^{(i)}}{\sum_{i=1}^{n} \mathbb{1}\{y^{(i)}=1\}}$$

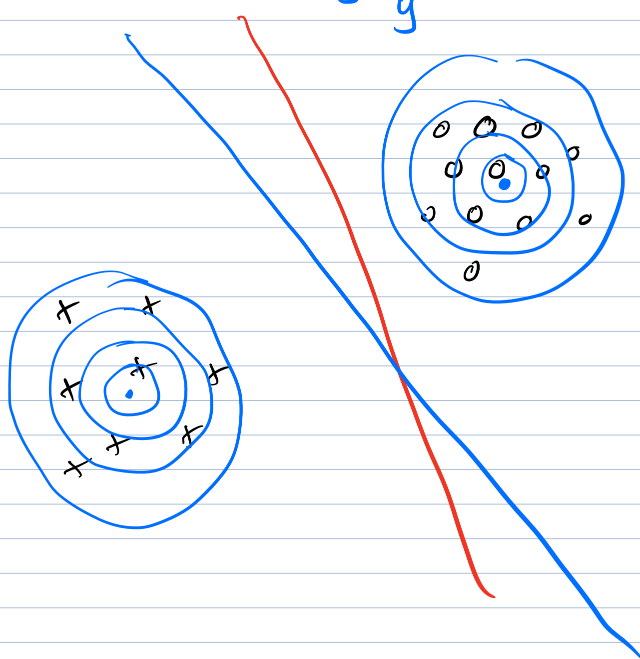
$$\Sigma = \frac{1}{n} \sum_{i=1}^{n} (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T$$

Prediction:

$$\arg\max_y p(y|x) = \arg\max_y \frac{p(x|y) \cdot p(y)}{p(x)}$$

e.g. $\min_z (z-2)^2 = 0 \qquad \arg\min_z (z-2)^2 = 2$

$$= \arg\max_y p(x|y) \cdot p(y)$$

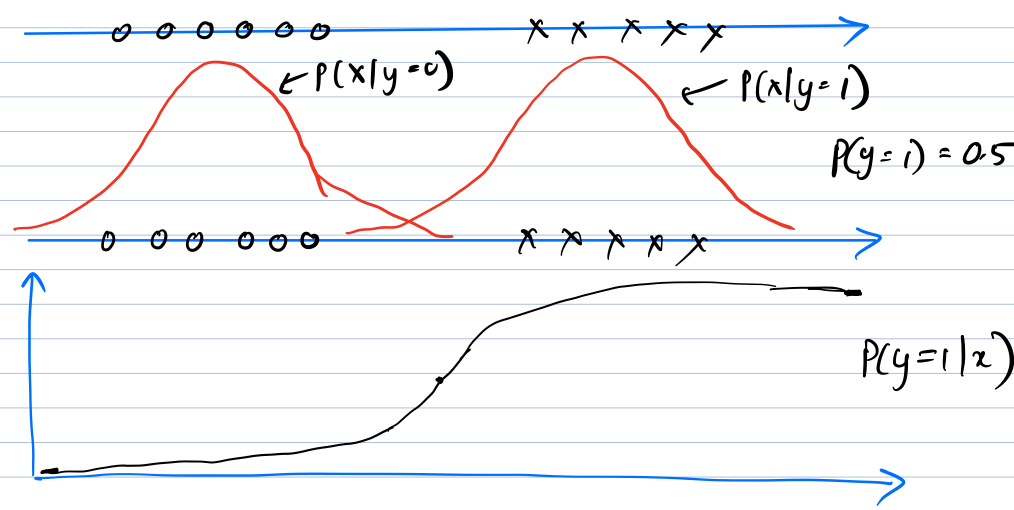


Comparison to Logistic Regression

for fixed $\phi, \mu_0, \mu_1, \Sigma$ lets

plot $p(y=1 \mid x; \phi, \mu_0, \mu_1, \Sigma)$ as f^n of x

$$\frac{p(x \mid y=1; \mu_1, \Sigma) \cdot p(y=1; \phi)}{p(x; \mu_0, \mu_1, \Sigma)}$$



Generative | Discriminative

GDA assumes

$x \mid y=0 \sim N(\mu_0, \Sigma)$

$x \mid y=1 \sim N(\mu_1, \Sigma)$

$y \sim Ber(\phi)$

$\Longrightarrow$

$\not\Longleftarrow$

Logistic Regression

$$P(y=1 \mid x) = \frac{1}{1+e^{-\theta^T x}}$$

"$x_0 = 1$"

Stronger assumption

Weaker Assumption

$x \mid y=1 \sim Poisson(\lambda_1)$

$x \mid y=0 \sim Poisson(\lambda_0)$

$y \sim Ber(\phi)$

$\Longrightarrow$ $P(y=1 \mid x)$ is logistic f^n

Naive Bayes

Running example: spam classifier

Feature vector x ?

$$x = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ \\ \\ \end{bmatrix} \begin{array}{l} \text{a} \\ \text{aardvark} \\ \text{aardwolf} \\ \vdots \\ \text{buy} \\ \text{CS229} \\ \vdots \\ \text{zymurgy} \end{array}$$

10,000 top 10K

$x \in \{0,1\}^d \quad d = 10,000$

$x_i = \mathbb{1}\{\text{word } i \text{ appears in email}\}$

Want to model $p(x|y)\ p(y)$

$2^{10,000}$ possible values of x

Assume x_i's are conditionally independent given y

$$p(x_1 \ldots x_{10,000}|y) = p(x_1|y) \cdot p(x_2|x_1,y) \cdot p(x_3|x_1,x_2,y) \cdots P(x_{10,000}| \ldots)$$

$$\overset{\text{assume}}{=} P(x_1|y) \cdot P(x_2|y) \cdot P(x_3|y) \ldots P(x_{10000}|y)$$

$$= \prod_{i=1}^{d} P(x_i|y)$$

Parameters: $\phi_{j|y=1} = P(x_j = 1 | y = 1)$ if it is a spam

$\phi_{j|y=0} = P(x_j = 1| y = 0)$ if it is not spam

$\phi_y = P(y=1)$ Pr(spam)