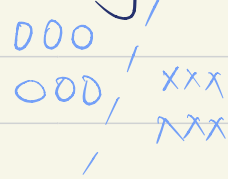
Lecture 5 GDA & Naive Bayes

1. Generative Learning Algorithms VS. Discriminative Comparison

Discriminative Learning Algorithms:

- Learns $p(y|x)$ or learns $h_\theta(x) = \{0, 1$ directly
- Learns a decision boundary
- Regressions, SVM



Generative Learning Algorithms:

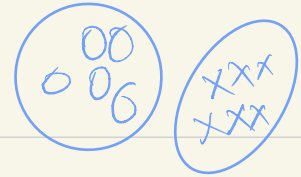
- Learns $P(x|y)$ and $P(y)$

feature, class, class prior

Using Bayesian Rules:

$$P(y=1|x) = \frac{P(x|y=1) \cdot P(y=1)}{P(x)}; \quad P(x) = P(x|y=1)P(y=1) + P(x|y=0)P(y=0)$$

- Learns probability distribution of data
- GDA ; Naive Bayes



2. Gaussian Discriminant Analysis (GDA)

Assumption:

- $x \in \mathbb{R}^n$ (drop $x_0 = 1$ convention)
- $P(x|y)$ is Gaussian

$\Rightarrow z \sim N(\vec{\mu}, \Sigma)$, $z \in \mathbb{R}^n$

($\vec{\mu} \in \mathbb{R}^n$, $\Sigma \in \mathbb{R}^{n \times n}$)

$$E[z] = \mu \; ; \; Cov(z) = E[(z-\mu)(z-\mu)^T] = E[z \cdot z^T] - (E[z])(E[z])^T$$

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

GDA Model:

$$P(x|y=0) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu_0)^T \Sigma^{-1}(x-\mu_0)\right)$$

$$P(x|y=1) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1)\right)$$

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

Parameters: $\mu_0, \mu_1, \Sigma, \phi (\in \mathbb{R} \in [0,1])$

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

Joint Likelihood: $L(\phi, \mu_0, \mu_1, \Sigma) = \prod_{i=1}^{m} P(x^{(i)}, y^{(i)}; \phi, \mu_0, \mu_1, \Sigma)$

$= \prod_{i=1}^{m} P(x^{(i)}|y^{(i)}) P(y^{(i)})$

<compare generative>: $L(\theta) = \prod_{i=1}^{n} P(y^{(i)}|x^{(i)}; \theta)$ ← conditional likelihood

Maximum Likelihood Estimation:

$$\max_{\phi, \mu_0, \mu_1, \Sigma} l(\phi, \mu_0, \mu_1, \Sigma)$$

$$\| \log(L(\phi, \mu_0, \mu_1, \Sigma))$$

$$\Rightarrow \phi = \frac{\sum_{i=1}^{m} (y^{(i)})}{m} = \frac{\sum_{i=1}^{m} I\{y^{(i)}=1\}}{m}$$ (the proportion of $y=1$)

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

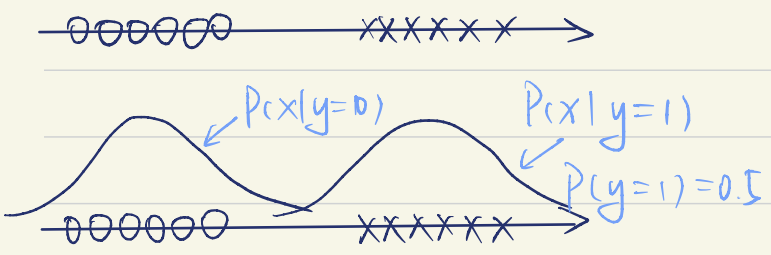
$$\mu_1 = \frac{\sum_{i=1}^{m} I\{y^{(i)}=1\} x^{(i)}}{\sum_{i=1}^{m} I\{y^{(i)}=1\}}$$

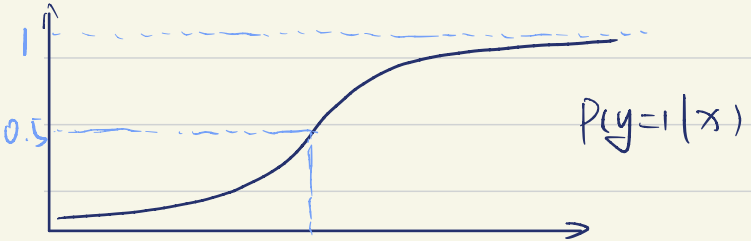
$$\Sigma = \frac{1}{m} \sum_{i=1}^{m} (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T$$

Prediction: $\underset{y}{\text{argmax}}\ P(y|x) = \underset{y}{\text{argmax}}\ \frac{P(x|y)P(y)}{P(x)}$

3. Compare GDA to Logistic Regression

for fixed $\phi, \mu_0, \mu_1, \Sigma$, plot $P(y=1 | x; \phi, \mu_0, \mu_1, \Sigma)$ as fn of x





(logistic Regression)

Generative

GDA assumes

$$\left.\begin{aligned} x|y=0 &\sim N(\mu_0, \Sigma) \\ x|y=1 &\sim N(\mu_1, \Sigma) \\ y &\sim Ber(\phi) \end{aligned}\right\} \Rightarrow P(y=1|x) = \frac{1}{1+e^{-\theta^T x}}$$

"$x_0 = 1$"

stronger assumption

$$\left.\begin{aligned} x|y=1 &\sim \text{Poisson}(\lambda_1) \\ x|y=0 &\sim \text{Poisson}(\lambda_0) \\ y &\sim Ber(\phi) \end{aligned}\right\} \Rightarrow P(y=1|x) \text{ is logistic fn}$$

Discriminative

Logistic Regression

weaker assumption

How to choose which method to use?

- Large data set – Logistic Regression (okay to tell the algorithm less)
- Computational Efficiency Matters – GDA

4. Naive Bayes (example: spam email classifier)

Feature Vector X?

$$X = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \leftrightarrow \begin{bmatrix} a \\ \text{aardvark} \\ \vdots \\ \text{zymurgy} \end{bmatrix} \quad \begin{matrix} \uparrow \\ 10,000 \\ \downarrow \end{matrix}$$

$X \in \{0,1\}^n$

$X_i = I\{\text{word } i \text{ appear in e-mail}\}$

Assumption: x_is are conditionally independent given y

$$P(x_1, \cdots, x_{10000} | y) = P(x_1|y) P(x_2|x_1, y) P(x_3|x_1, x_2, y) \cdots P(x_{10000}|\cdots)$$
$$\overset{\text{assume}}{=} P(x_1|y) P(x_2|y) P(x_3|y) \cdots P(x_{10000}|y)$$
$$= \prod_{i=1}^{n} P(x_i|y)$$

Parameters:

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

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

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

Joint Likelihood: $L(\phi_y, \phi_{j|y}) = \prod_{i=1}^{m} P(x^{(i)}, y^{(i)}, \phi_y, \phi_{j|y})$

MLE: $\phi_y = \frac{\sum_{i=1}^{m} I\{y^{(i)} = 1\}}{m}$ $\quad$ $\phi_{j|y=1} = \frac{\sum_{i=1}^{m} I\{x_j^{(i)} = 1, y^{(i)} = 1\}}{\sum_{i=1}^{m} I\{y^{(i)} = 1\}}$