Generative Learning Algorithms

Build models for what each class looks like and classify new points using these models

Formally:

Learns $p(x|y)$ and $p(y)$

features class class prior

(given a tumor is malignant/benign, what do its features look like?)

(what is the probability of any tumor being malignant/benign?)

Bayes Rule

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

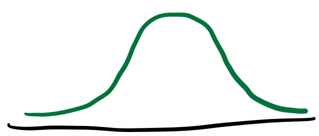
$$= \underbrace{p(x|y=1)p(y=1) + p(x|y=0)p(y=0)}_{\text{learn all these terms}}$$

Gaussian Discriminant Analysis

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

Assume $p(x|y)$ is Gaussian

Multivariate Gaussian



$$z \sim N(\underset{\mathbb{R}^d}{\vec{\mu}}, \underset{\mathbb{R}^{d\times d}}{\Sigma})$$

$z = (z_1, z_2, \dots z_d) \in \mathbb{R}^d$

$$E[z] = \vec{\mu}$$

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

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

GDA model

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

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

Parameters: $\overbrace{\mu_0, \mu_1}^{\mathbb{R}^d}, \overset{\mathbb{R}^{d\times d}}{\Sigma}, \overset{\mathbb{R}\,\in [0,1]}{\phi}$

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

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

Joint likelihood

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

Maximum Likelihood Estimation

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

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

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

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

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

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

Prediction

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

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

Comparison w/ LR

GDA assumes

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

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

$y \sim Ber(\phi)$

LR assumes

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

Stronger assumption

$x|y=1 \sim Poi(\lambda_1)$

$x|y=0 \sim Poi(\lambda_0)$

$y \sim Ber(\phi)$

Weaker assumption

$p(y=1|x)$ vs logistic fn

Naive Bayes

Feature vector x (ex. spam classification)

$$x = \begin{array}{c} \left[\begin{array}{c} 1 \\ 0 \\ 0 \\ \vdots \\ \\ \end{array}\right] \end{array} \begin{array}{l} \text{a} \\ \text{aardvark} \\ \\ \vdots \\ \text{zymurgy} \end{array}$$

$\uparrow$ top-d words $\downarrow$

$$x \in \{0,1\}^d$$

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

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

2^d possible values of x

Assume x_i's are conditionally independent given y

$$\Rightarrow p(x_1 \dots x_d | y) = p(x_1|y)\, p(x_2|x_1,y) \cdots p(x_d|x_{d-1} \dots x_1, y)$$

$$\overset{\text{assume}}{=} p(x_1|y)\, p(x_2|y) \cdots p(x_d|y)$$

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

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

$\phi_y = p(y)$ Pr(spam)