CS 307 Lecture 16

Classification : $S = \{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$, $y^{(i)} \in \{0,1\}$

<u>Discriminative Learning</u> : The shape of the function that will determine the output on new examples

- Determine function by finding the one that minimizes loss function.
- Determine $p(y \mid x; \theta)$ by maximum likelihood estimation

Prediction: ~~One~~ On (x),

compute $p(y=1 \mid x; \theta)$, $p(y=0 \mid x; \theta)$

(MAP) <u>Maximum a-Posterior estimation</u>:

$$\text{Output} = \operatorname{argmax}\left[\, p(y=1 \mid x;\theta),\ p[y=0 \mid x;\theta]\right)$$

<u>Generative Learning</u>: Model $\underbrace{\boxed{p(x \mid y)}}_{\text{depend on parameters}}$, $\boxed{p(y)}$

$$p(\{(x^{(i)}, y^{(i)}\}; \theta) = \prod_{i=1}^{n} p(x^{(i)}, y^{(i)}; \theta)$$

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

Pick parameters that maximize likelihood.

For prediction.

Bayes Rule: $p(y=1 \mid x) = \dfrac{p(x, y=1)}{p(x)}$

$$= \frac{p(y=1) \cdot p(x \mid y=1)}{p(y=0)\cdot p(x \mid y=0) + p(y=1)\cdot p(x \mid y=1)}$$

Classification by MAP rule.

Univariate Gaussian Distribution

$$p_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

probability density fn.

$$\sigma^2 = E[(X-\mu)^2]$$

Multivariate Gaussian Distribution:

x is a vector of $\mathbb{R}$-valued random variables. (x - vector of values).

$$\mu \in \mathbb{R}^d, \quad \underbrace{\Sigma}_{\text{Covariance matrix}} \in \mathbb{R}^{d\times d}$$

$$\to E(x) = \int x\, p(x)\, dx \qquad = E[(X-\mu)(X-\mu)^T]$$

$$p_{\mu,\Sigma}(x) = \frac{1}{(2\pi)^{d/2}\underbrace{|\Sigma|}_{\text{determinant } \Sigma}^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right]$$

Gaussian Discriminant Analysis (GDA):

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

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

$$p(x\mid y=0) \sim N(\mu_0, \Sigma) \qquad p(x\mid y=1) \sim N(\mu_1, \Sigma)$$

$$p(y) = \begin{cases} \phi & y=1 \\ 1-\phi & y=0 \end{cases} \Rightarrow p(y) = \phi^y(1-\phi)^{1-y}.$$

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

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

$$\max\ \ell(\mu_0,\mu_1,\Sigma,\phi) = \max\ \ln L(\mu_0,\mu_1,\Sigma,\phi)$$

$$\nabla_{\mu_0}\ell(\mu_0,\mu_1,\Sigma,\phi) = 0, \quad \nabla_{\mu_1}\ell(\mu_0,\mu_1,\Sigma,\phi) = 0$$

$$\nabla_{\Sigma}\ell(\mu_0,\mu_1,\Sigma,\phi) = 0, \quad \nabla_{\phi}\ell(\mu_0,\mu_1,\Sigma,\phi) = 0$$

Values of $\mu_0, \mu_1, \Sigma, \phi$ that maximize likelihood

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

$$\mu_0 = \frac{\sum_{i=1}^{n} x^{(i)} 1(y^{(i)}=0)}{\sum_{i=1}^{n} 1(y^{(i)}=0)} \qquad \mu_1 = \frac{\sum_{i=1}^{n} x^{(i)} 1[y^{(i)}=1]}{\sum_{i=1}^{n} 1[y^{(i)}=1]}$$

$$\Sigma = \frac{1}{n^2} \sum_{i=1}^{n} \left(x^{(i)} - \mu_{y^{(i)}}\right)\left(x^{(i)} - \mu_{y^{(i)}}\right)^T$$

__Prediction__: Compute $p(y=1|x)$ and $p(y=0|x)$

Output 1 provided

$$\ln \left[\frac{p(x|y=1) \cdot p(y=1)}{\cancel{p(x)}} \cdot \frac{\cancel{p(x)}}{p(x|y=0) \cdot p(y=0)} \right] > 0$$

$$\ln[p(x|y=1)] + \ln(p(y=1)) - \ln[p(x|y=0)] - \ln[p(y=0)] > 0$$

$$\underbrace{-\frac{1}{2}[(x-\mu_0)^T \Sigma^{-1} (x-\mu_0)] + \ln\phi \quad + \frac{1}{2}[(x-\mu_1)^T \Sigma^{-1} (x-\mu_1)] - \ln(1-\phi)}_{} > 0$$

$$\theta^T x + \theta_0 > 0 \qquad \Sigma = \Sigma^T$$

depend on $\mu_0, \mu_1, \Sigma, \phi$. $\qquad \Sigma^{-1} = (\Sigma^{-1})^T$

$$p(y|x) = \frac{1}{1 + e^{-(\theta^T x + \theta_0)}} \quad \Big\} \text{ logistic fn.}$$

Learning process is more expensive for logistic regression.

- If $p(x|y)$ = distribution $\in$ Exponential family

  then $p(y|x)$ is logistic fn.

Logistic regression is more robust to modeling assumptions