Generative Learning algorithms

Discriminative vs generative

discriminative: learn $p(y|x)$ directly or mapping directly from X to y.

generative: $p(y|x) = \frac{p(x|y)p(y)}{p(x)} \Rightarrow$ model $p(x|y)$

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

Gaussian discriminative analysis (GDA)

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

Where, $\mu \in R^d$, $\Sigma \in R^{n\times n}$ (covariance matrix).

假设特征x是连续随机变量，则GDA模型为：

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

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

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

∴ log likelihood is

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

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

$$= \sum_{i=1}^{n} \log p(x^{(i)} | y^{(i)}; \mu_0, \mu_1, \Sigma) + \sum_{i=1}^{n} \log p(y^{(i)}; \phi)$$

$$\therefore \frac{\partial \ell}{\partial \phi} = \frac{\partial}{\partial \phi} \sum_{i=1}^{n} \log p(y^{(i)}; \phi)$$

$$= \frac{\partial}{\partial \phi} \sum_{i=1}^{n} \log \left[\phi^{y^{(i)}} (1-\phi)^{1-y^{(i)}} \right]$$

$$= \frac{\partial}{\partial \phi} \sum_{i=1}^{n} \left[y^{(i)} \log \phi + (1-y^{(i)}) \log(1-\phi) \right]$$

$$= \frac{\partial}{\partial \phi} \sum_{i=1}^{n} \left[y^{(i)} \log \phi + \log(1-\phi) - y^{(i)} \log(1-\phi) \right]$$

$$= \sum_{i=1}^{n} \left(y^{(i)} \frac{1}{\phi} + \frac{1}{1-\phi}(-1) - y^{(i)} \frac{1}{1-\phi}(-1) \right)$$

$$= \sum_{i=1}^{n} \left(y^{(i)}(1-\phi) - \phi + y^{(i)}\phi \right) = \sum_{i=1}^{n} \left(y^{(i)} - \phi \right) = 0$$

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

$$\therefore \frac{\partial \ell}{\partial \mu_0} = \frac{\partial}{\partial \mu_0} \log \exp\left[-\frac{1}{2} (x-\mu_0)^T \Sigma^{-1} (x-\mu_0) \right]$$

$$= -\frac{1}{2} \frac{\partial}{\partial \mu_0} \sum_{i=1}^{n} (x^{(i)}-\mu_0)^T \Sigma^{-1} (x^{(i)}-\mu_0)$$

$$= -\frac{1}{2} \sum_{i=1}^{n} \left[\Sigma^{-1} (x^{(i)}-\mu_0)(-1) + (\Sigma^{-1})^T (x^{(i)}-\mu_0) \right]$$

$$= -\frac{1}{2} \left((\Sigma^{-1})^T - \Sigma^{-1} \right) (x^{(i)}-\mu_0) = 0 \quad \Rightarrow \mu_0 = \frac{\sum_{i=1}^{n} I\{y^{(i)}=0\}\, x^{(i)}}{\sum_{i=1}^{n} I\{y^{(i)}=0\}}$$

同理: $\frac{\partial \ell}{\partial \mu_1} = 0 \Rightarrow \mu_1 = \frac{\sum_{i=1}^{n} I\{y^{(i)}=1\}\, x^{(i)}}{\sum_{i=1}^{n} I\{y^{(i)}=1\}}$

设 $S = \Sigma^{-1}$

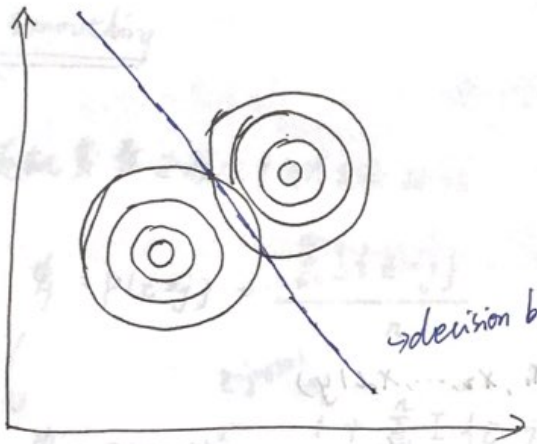
$$\therefore \frac{\partial L}{\partial \Sigma} = \frac{\partial}{\partial \Sigma} - \frac{1}{2} \sum_{i=1}^{n} \left[\log |S|^{-1} + (x^{(i)} - \mu_{y^{(i)}})^T S (x^{(i)} - \mu_{y^{(i)}}) \right]$$

$$= \frac{\partial}{\partial \Sigma} - \frac{1}{2} \sum_{i=1}^{n} \left[-\log |S| + (x^{(i)} - \mu_{y^{(i)}})^T S (x^{(i)} - \mu_{y^{(i)}}) \right]$$

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

$$= \frac{1}{2} \sum_{i=1}^{n} \Sigma - \frac{1}{2} \sum_{i=1}^{n} (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T = 0$$

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



GDA and logistic regression

当 $p(x|y)$ 是多元高斯分布时，$p(y|x)$ 满足逻辑回归的形式.

$$p(y=1|x;\phi,\Sigma,\mu_0,\mu_1)=\frac{1}{1+\exp(-\theta^T x)}$$

⇒ 当 $p(y|x)$ 满足高斯分布时 ~~[illegible]~~ GDA 的结果 LR。

Naive Bayes

特征是离散的.

Naive Bayes Assumption.

$$p(x_1,\dots,x_d|y)=$$

$$=p(x_1|y)\,p(x_1,x_2|y)\cdots p(x_1,x_2,\dots,x_d|y)$$

$$=p(x_1|y)\,p(x_2|y)\cdots p(x_d|y)=\prod_{j=1}^{d}p(x_j|y)$$

$$L(\phi_y,\phi_{j|y=0},\phi_{j|y=1})=\prod_{i=1}^{n}p(x^{(i)}|y^{(i)})\,p(y^{(i)})$$

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

$$\Rightarrow \phi_{j|y=1}=\frac{\sum_{i=1}^{n}I\{x_j^{(i)}=1\wedge y^{(i)}=1\}}{\sum_{i=1}^{n}I\{y^{(i)}=1\}},\quad \phi_{j|y=0}=\frac{\sum_{i=1}^{n}I\{x_j^{(i)}=1\wedge y^{(i)}=0\}}{\sum_{i=1}^{n}I\{y^{(i)}=0\}}$$

$$\therefore \phi_y=\frac{\sum_{i=1}^{n}I\{y^{(i)}=1\}}{n}$$

预测：

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

$$\propto P(x|y=1)\,P(y=1)$$

$$= \prod_{j=1}^{d} P(x_j|y=1) \cdot P(y=1)$$

假设输入的特征 x_j 可以取 k 种不同的值，则 $P(x_j|y)$ 在该建模为多项式分布。且各为为连续值，需要进行离散化操作。

Laplace smoothing

假设随机变量 z 服从多项式分布，从 $1, \dots, k$ 中取值。给定 n 个训练样本则

$$\phi_j = P(z=j) = \frac{\sum_{i=1}^{n} I\{z=j\}}{n}$$

smoothing ↓ （$z=j$ 时为 1）

$$\phi_j' = P(z=j)' = \frac{1 + \sum_{i=1}^{n} I\{z=j\}}{k+n}$$

（k：此特征可以取值）

text classification with nb

假设 d_i 表示第 i 个文档，c 表示类别且有 K 种取值。

1) Bernoulli model

文档使用 one hot 方式表示，

$$p(d_i, c) = p(d_i|c)\,p(c) = \prod_{j=1}^{V} p(d_{ij}|c)\,p(c)$$

$$p(c) = \frac{n_k}{n} \quad \begin{cases} \rightarrow \text{属于类别 } k \text{ 的文档数} \\ \rightarrow \text{总文档数} \end{cases}$$

$$p(d_{ij}|c) = \frac{n_k(w_j)}{n_k} \rightarrow \text{属于类别 } k \text{ 且包含单词 } w_j \text{ 的文档数}$$

Laplace smoothing $\rightarrow$ $\hat{p}(d_{ij}|c) = \dfrac{n_k(w_j)+1}{n_k+2}$

2) Multinomial model

文档使用 multi-hot (term-frequency) 方式表示。

$$p(d_i, c) = p(d_i|c)\,p(c) = \prod_{j=1}^{V} p(d_{ij}|c)\,p(c)$$

$$p(c) = \frac{\cancel{n_k(w_j)}}{\cancel{n_k}} \quad \frac{n_k}{n}$$

$$p(d_{ij}|c) = \frac{nw_k(w_j)}{nw_k} \rightarrow \text{属于类别 } k \text{ 的文档中 } w_j \text{ 出现的次数和}$$

$\rightarrow$ 属于类别 k 的文档中单词出现的总次数之和

Laplace smoothing $\downarrow$

$$\hat{p}(d_{ij}|c) = \frac{nw_k(w_j)+1}{nw_k+V}$$