

Naive Bayes and EM Algorithm

Theorem.

Bayes' Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- If X and Y are both continuous, then $f_{X|Y=y}(x) = \frac{f_{Y|X=x}(y)f_X(x)}{f_Y(y)}$

Law of total probability

If $\{B_n : n = 1, 2, 3, \dots\}$ is a finite or countably infinite partition of a sample space.

$$P(A) = \sum_n P(A | B_n)P(B_n)$$

Warm up

- In linear regression and logistic regression, x and y are linked through (deterministic) hypothesis function
- Given a set of training data $\mathcal{D} = \{x^{(i)}, y^{(i)}\}_{i=1, \dots, m}$, $P(\mathcal{D}) = \prod_{i=1}^m p_{X|Y}(x^{(i)}|y^{(i)})p_Y(y^{(i)})$

Gaussian Distribution

- Normal Distribution $p(x; \mu, \sigma) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right)$
- Multivariate normal distribution $p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right)$
- where $\mu \in \mathbb{R}^n$, $\Sigma \in \mathbb{R}^{n \times n}$ is symmetric and positive semidefinite

Gaussian Discriminant Analysis (GDA)

- $Y \sim \text{Bernoulli}(\psi)$
- $p_{X|Y}(x | 0) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_0)^T \Sigma^{-1}(x - \mu_0)\right)$
- $p_{X|Y}(x | 1) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1)\right)$

$$\ell(\psi, \mu_0, \mu_1, \Sigma) = \sum_{i=1}^m \log p_{X|Y}(x^{(i)} | y^{(i)}; \mu_0, \mu_1, \Sigma) + \sum_{i=1}^m \log p_Y(y^{(i)}; \psi)$$

- Maximizing $\ell(\psi, \mu_0, \mu_1, \Sigma)$ to get the solutions:

$$\begin{aligned}
\psi &= \frac{1}{m} \sum_{i=1}^m 1 \{y^{(i)} = 1\} \\
\mu_0 &= \sum_{i=1}^m 1 \{y^{(i)} = 0\} x^{(i)} / \sum_{i=1}^m 1 \{y^{(i)} = 0\} \\
\mu_1 &= \sum_{i=1}^m 1 \{y^{(i)} = 1\} x^{(i)} / \sum_{i=1}^m 1 \{y^{(i)} = 1\} \\
\Sigma &= \frac{1}{m} \sum_{i=1}^m (x^{(i)} - \mu_{y^{(i)}}) (x^{(i)} - \mu_{y^{(i)}})^T
\end{aligned}$$

- Given a test data sample x , we can calculate

$$\begin{aligned}
p_{Y|X}(y = 1 | x) &= \frac{p_{X|Y}(x | 1)p_Y(1)}{p_X(x)} \\
&= \frac{p_{X|Y}(x | 1)p_Y(1)}{p_{X|Y}(x | 1)p_Y(1) + p_{X|Y}(x | 0)p_Y(0)} \\
&= \frac{1}{1 + \frac{p_{X|Y}(x|0)p_Y(0)}{p_{X|Y}(x|1)p_Y(1)}}
\end{aligned}$$

Naive Bayes

Assumption.

For $\forall j \neq j'$ X_j and $X_{j'}$ are conditionally independent given Y , i.e.,
 $P(Y = y, X_1 = x_1, \dots, X_n = x_n) = p(y) \prod_{j=1}^n p_j(x_j | y)$

- MLE $\ell(\Omega) = \sum_{i=1}^m \log p(y^{(i)}) + \sum_{i=1}^m \sum_{j=1}^n \log p_j(x_j^{(i)} | y^{(i)})$

Theorem.

$$\begin{aligned}
p(y) &= \frac{\text{count}(y)}{m}, & p_j(x | y) &= \frac{\text{count}_j(x|y)}{\text{count}(y)} \\
\text{count}(y) &= \sum_{i=1}^m 1(y^{(i)} = y), & \text{count}_j(x | y) &= \sum_{i=1}^m 1(y^{(i)} = y \wedge x_j^{(i)} = x), \\
&\forall y = 0, 1, & &\forall x = 0, 1.
\end{aligned}$$

Classification by NB

$$\begin{aligned}
\therefore P(Y = y | X_1 = \tilde{x}_1, \dots, X_n = \tilde{x}_n) &= \frac{P(X_1 = \tilde{x}_1, \dots, X_n = \tilde{x}_n | Y = y) P(Y = y)}{P(X_1 = \tilde{x}_1, \dots, X_n = \tilde{x}_n)} \\
\therefore \arg \max_{y \in \{0,1\}} (p(y) \prod_{j=1}^n p_j(\tilde{x}_j | y))
\end{aligned}$$

Laplace Smoothing

- There may exist some feature, e.g., X_{j^*} , such that $X_{j^*}=1$ for some x^* may never happen in the training data. (i.e., $p_{j^*}(x_{j^*} = 1 | y) = \frac{\sum_{i=1}^m 1(y^{(i)}=y \wedge x_{j^*}^{(i)}=1)}{\sum_{i=1}^m 1(y^{(i)}=y)} = 0, \forall y = 0, 1$)

- $\therefore p(y | x) = \frac{p(y) \prod_{j=1}^n p_j(x_j | y)}{\sum_y \prod_{j=1}^n p_j(x_j | y) p(y)} = \frac{0}{0}, \forall y = 0, 1$

Laplace Smoothing

$$p(y) = \frac{\sum_{i=1}^m \mathbf{1}(y^{(i)} = y) + 1}{m + k}$$

$$p_j(x | y) = \frac{\sum_{i=1}^m \mathbf{1}(y^{(i)} = y \wedge x_j^{(i)} = x) + 1}{\sum_{i=1}^m \mathbf{1}(y^{(i)} = y) + v_j}$$

where k is number of the possible values of y ($k = 2$ in our case), and v_j is the number of the possible values of the j -th feature ($v_j = 2$ for $\forall j = 1, \dots, n$ in our case)

Multinomial Distribution

Assumption

Each training sample involves a different number of features

$$x^{(i)} = [x_1^{(i)}, x_2^{(i)}, \dots, x_{n_i}^{(i)}]^\top$$

The j -th feature of $x^{(i)}$ takes a finite set of values, $x_j^{(i)} \in \{1, 2, \dots, v\}$

- For example, $x_j^{(i)}$ indicates the j -th word in the email.
- Let $p(t | y) = P(X_j = t | Y = y)$

$$P(Y = y^{(i)}) = p(y^{(i)})$$

$$P(X = x^{(i)} | Y = y^{(i)}) = \prod_{j=1}^{n_i} p(x_j^{(i)} | y^{(i)})$$

Problem.

$$\max \quad \ell(\Omega) = \log p(y^{(i)}) \prod_{i=1}^m p(x^{(i)} | y^{(i)}) = \sum_{i=1}^m \sum_{j=1}^{n_i} \log p(x_j^{(i)} | y^{(i)}) + \sum_{i=1}^m \log p(y^{(i)})$$

$$\begin{aligned} \text{s.t.} \quad & \sum_{y \in \{0,1\}} p(y) = 1, \\ & \sum_{t=1}^v p(t | y) = 1, \forall y = 0, 1 \\ & p(y) \geq 0, \forall y = 0, 1 \\ & p(t | y) \geq 0, \forall t = 1, \dots, v, \forall y = 0, 1 \end{aligned}$$

- Solution

$$p(t | y) = \frac{\sum_{i=1}^m \mathbf{1}(y^{(i)} = y) \text{count}^{(i)}(t)}{\sum_{i=1}^m \mathbf{1}(y^{(i)} = y) \sum_{t=1}^v \text{count}^{(i)}(t)}$$

$$p(y) = \frac{\sum_{i=1}^m \mathbf{1}(y^{(i)} = y)}{m}$$

$$\text{where } \text{count}^{(i)}(t) = \sum_{j=1}^{n_i} \mathbf{1}(x_j^{(i)} = t)$$

- Laplace smoothing

$$\psi(y) = \frac{\sum_{i=1}^m 1(y^{(i)}=y)+1}{m+k}$$

$$\psi(t | y) = \frac{\sum_{i=1}^m 1(y^{(i)}=y) \text{count}^{(i)}(t)+1}{\sum_{i=1}^m 1(y^{(i)}=y) \sum_{t=1}^v \text{count}^{(i)}(t)+v}$$

Expectation Maximization (EM) Algorithm

Def. Latent Variable.

$$\ell(\theta) = \log \prod_{i=1}^m p(x^{(i)}; \theta)$$

$$= \sum_{i=1}^m \log \sum_{z^{(i)} \in \Omega} p(x^{(i)}, z^{(i)}; \theta)$$

where $z^{(i)} \in \Omega$ is so-called latent variable.

- Basic idea of EM algorithm
 - Repeatedly construct a lower-bound on ℓ (E-step)
 - * E-Step estimates the parameters by observing the data and the existing model, and then use this estimated parameter value to calculate the expected value of the likelihood function.
 - Then optimize that lower-bound (M-step)
 - * M-step finds the corresponding parameters when the likelihood function is maximized. Since the algorithm guarantees that the likelihood function will increase after each iteration, the function will eventually converge.
- Let Q_i denotes the distribution of z for i -th sample. Thus, $\sum_{z \in \Omega} Q_i(z) = 1, Q_i(z) \geq 0$

$$\ell(\theta) = \sum_{i=1}^m \log \sum_{z^{(i)} \in \Omega} Q_i(z^{(i)}) \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}$$

$$= \sum_{i=1}^m \log E \left[\frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} \right]$$

Theorem. Jensen's Inequality.

Assume f be a concave function.

$$f(E[X]) \geq E(f(X))$$

- Since $\log(\cdot)$ is a concave function, according to Jensen's inequality, we have

$$\therefore \log \left(E \left[\frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} \right] \right) \geq E \left[\log \left(\frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} \right) \right]$$

$$\therefore \ell(\theta) \geq \sum_{i=1}^m \sum_{z^{(i)} \in \Omega} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}$$

- Tighten the lower bound, the equality holds when $\frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} = c$, where c is a constant.
- Therefore, $\sum_{z^{(i)} \in \Omega} p(x^{(i)}, z^{(i)}; \theta) = c \sum_{z^{(i)} \in \Omega} Q_i(z) = c$

$$\begin{aligned}
Q_i(z^{(i)}) &= \frac{p(x^{(i)}, z^{(i)}; \theta)}{c} \\
&= \frac{p(x^{(i)}, z^{(i)}; \theta)}{\sum_{z^{(i)} \in \Omega} p(x^{(i)}, z^{(i)}; \theta)} \\
&= \frac{p(x^{(i)}, z^{(i)}; \theta)}{p(x^{(i)}; \theta)} \\
&= p(z^{(i)} | x^{(i)}; \theta)
\end{aligned}$$

Algorithm.

1. (E-step) For each i , set $Q_i(z^{(i)}) := p(z^{(i)} | x^{(i)}; \theta)$
2. (M-step) set $\theta := \arg \max_{\theta} \sum_i \sum_{z^{(i)} \in \Omega} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}$

EM in NB

Naive Bayes with Missing Labels

- When labels are given $\ell(\theta) = \log p(x, y) = \sum_{i=1}^m \log [p(y^{(i)}) \prod_{j=1}^n p_j(x_j^{(i)} | y^{(i)})]$
- When labels are missed $\ell(\theta) = \log p(x) = \sum_{i=1}^m \log \sum_{y=1}^k [p(y) \prod_{j=1}^n p_j(x_j^{(i)} | y)]$

Applying EM to NB

- (E-step) For each $i = 1, \dots, m$ and $y = 1, \dots, k$ set
 - Relabel y by $Q_i(y)$.

$$Q_i(y) = p(y^{(i)} = y | x^{(i)}) = \frac{p(y) \prod_{j=1}^n p_j(x_j^{(i)} | y)}{\sum_{y'=1}^k p(y') \prod_{j=1}^n p_j(x_j^{(i)} | y')}$$

- (M-step) Update the parameters (solved by Lagrange multiplier).
 - Use $\sum_{i=1}^m Q_i(y)$ to substitute the $count(y)$

$$\begin{aligned}
p(y) &= \frac{1}{m} \sum_{i=1}^m Q_i(y), \quad \forall y \\
p_j(x | y) &= \frac{\sum_{i: x_j^{(i)} = x} Q_i(y)}{\sum_{i=1}^m Q_i(y)}, \quad \forall x, y
\end{aligned}$$