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

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

Warm up

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

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 postitive semidefinite

~~Gaussian Discriminant Analysis (GDA)~~

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

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

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

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

- Given a test data sample x, we can calculate

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

Naive Bayes

Assumption.

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

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

Theorem.

$$
p(y)=\frac{count(y)}{m}, p_j(x \mid y)=\frac{count_j(x \mid y)}{count(y)} count(y)=\sum_{i=1}^{m} 1(y^{(i)}=y),\ count_j(x \mid y)=\sum_{i=1}^{m} 1(y^{(i)}=y \wedge x_j^{(i)} =
$$

Classification by NB

$$
\because P(Y=y \mid X_1=\tilde{x_1}, \cdots, X_n=\tilde{x_n})=\frac{P(X_1=\tilde{x_1}, \cdots, X_n=\tilde{x_n} \mid Y=y)P(Y=y)}{P(X_1=\tilde{x_1}, \cdots, X_n=\tilde{x_n})} \therefore \arg\max_{y\in\{0,1\}}\left(p(y)\prod_{j=1}^{n} p_j\left(\tilde{x}_j \mid y\right.\right.
$$

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., $p_{j^*}\left(x_{j^*}=1 \mid y\right)=\frac{\sum_{i=1}^{m} \mathtt{1}\left(y^{(i)}=y \wedge x_{j^*}^{(i)}=1\right)}{\sum_{i=1}^{m} \mathtt{1}(y^{(i)}=y)}=0, \forall y=0,1$)

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

  Laplace Smoothing

$$
\begin{aligned}
p(y) &= \frac{\sum_{i=1}^m \mathtt{1}\left(y^{(i)}=y\right)+1}{m+k} \\
p_j(x \mid y) &= \frac{\sum_{i=1}^m \mathtt{1}\left(y^{(i)}=y \wedge x_j^{(i)}=x\right)+1}{\sum_{i=1}^m \mathtt{1}\left(y^{(i)}=y\right)+v_j}
\end{aligned}
$$

  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 $\left(v_j=2\right.$ for $\forall j=1, \cdots, n$ in our case)

Multinomial Distribution

  Assumption

  Each training sample involves a different number of features

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

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

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

$$
\begin{aligned}
& P\left(Y=y^{(i)}\right)=p\left(y^{(i)}\right) \\
& P\left(X=x^{(i)} \mid Y=y^{(i)}\right)=\prod_{j=1}^{n_i} p\left(x_j^{(i)} \mid y^{(i)}\right)
\end{aligned}
$$

  Problem.

$$
\begin{array}{ll}
\max & \ell(\Omega)=\log p\left(y^{(i)}\right) \prod_{i=1}^m p(x^{(i)} \mid y^{(i)})=\sum_{i=1}^m \sum_{j=1}^{n_i} \log p(x_j^{(i)} \mid y^{(i)})+\sum_{i=1}^m \log p\left(y^{(i)}\right) \\
\text{s.t.} & \sum_{y \in\{0,1\}} p(y)=1, \\
& \sum_{t=1}^v p(t \mid y)=1, \forall y=0,1 \\
& p(y) \geq 0, \forall y=0,1 \\
& p(t \mid y) \geq 0, \forall t=1, \cdots, v, \forall y=0,1
\end{array}
$$

- Solution

$$
\begin{aligned}
& p(t \mid y)=\frac{\sum_{i=1}^m \mathtt{1}\left(y^{(i)}=y\right) \text{ count }^{(i)}(t)}{\sum_{i=1}^m \mathtt{1}\left(y^{(i)}=y\right) \sum_{t=1}^v \text{ count }^{(i)}(t)} \\
& p(y)=\frac{\sum_{i=1}^m \mathtt{1}\left(y^{(i)}=y\right)}{m} \\
& \text{where count }^{(i)}(t)=\sum_{j=1}^{n_i} \mathtt{1}\left(x_j^{(i)}=t\right)
\end{aligned}
$$

- Laplace smoothing

$$\psi(y) = \frac{\sum_{i=1}^{m} \mathbb{1}\left(y^{(i)} = y\right) + 1}{m + k}$$
$$\psi(t \mid y) = \frac{\sum_{i=1}^{m} \mathbb{1}\left(y^{(i)} = y\right) \mathrm{count}^{(i)}(t) + 1}{\sum_{i=1}^{m} \mathbb{1}\left(y^{(i)} = y\right) \sum_{t=1}^{v} \mathrm{count}^{(i)}(t) + v}$$

Expectation Maximization (EM) Algorithm

Def. Latent Variable.

$$\begin{aligned} \ell(\theta) &= \log \prod_{i=1}^{m} p(x^{(i)}; \theta) \\ &= \sum_{i=1}^{m} \log \sum_{z^{(i)} \in \Omega} p(x^{(i)}, z^{(i)}; \theta) \end{aligned}$$

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, $\sum_{z \in \Omega} Q_i(z) = 1, Q_i(z) \geq 0$

$$\begin{aligned} \ell(\theta) &= \sum_{i=1}^{m} \log \sum_{z^{(i)} \in \Omega} Q_i\left(z^{(i)}\right) \frac{p\left(x^{(i)}, z^{(i)}; \theta\right)}{Q_i\left(z^{(i)}\right)} \\ &= \sum_{i=1}^{m} \log E\left[\frac{p\left(x^{(i)}, z^{(i)}; \theta\right)}{Q_i\left(z^{(i)}\right)}\right] \end{aligned}$$

Thereom. Jesen'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

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

- 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, $\sum_{z^{(i)}\in\Omega} p(x^{(i)}, z^{(i)};\theta) = c\sum_{z^{(i)}\in\Omega} Q_i(z) = c$

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

Algorithm.

1. (E-step) For each i, set $Q_i(z^{(i)}) := p(z^{(i)} \mid x^{(i)};\theta)$
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 $\ell(\theta) = \log p(x,y) = \sum_{i=1}^{m} \log\left[\, p(y^{(i)}) \prod_{j=1}^{n} p_j(x_j^{(i)} \mid y^{(i)})\right]$
- When labels are missed $\ell(\theta) = \log p(x) = \sum_{i=1}^{m} \log\sum_{y=1}^{k}\left[p(y)\prod_{j=1}^{n} p_j(x_j^{(i)} \mid y)\right]$

Applying EM to NB

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

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

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

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