Lecture Notes on Gaussian Discriminant Analysis, Naive Bayes and EM Algorithm

Feng Li
fli@sdu.edu.cn
Shandong University, China

1 Gaussian Discriminant Analysis

In some application scenario, given a test data sample x, we may like to figure out the confidence in which x can be categorized into class $y = 0$ (or $y = 1$). In particular, the output of our leaning algorithm would be $p(y = 1 \mid x)$ and $p(y = 0 \mid x)$, i.e., the probability that x belongs to the class $y = 1$ (or $y = 0$). In fact, we can further determine the label of x by

$$y = \begin{cases} 0, & \text{if } p(y = 0 \mid x) \geq p(y = 1 \mid x) \\ 1, & \text{if } p(y = 0 \mid x) < p(y = 1 \mid x) \end{cases} \tag{1}$$

In *Gaussian Discriminate Analysis* (GDA) model, we have the following assumptions:

- $y \sim \mathit{Bernoulli}(\psi)$
- $x \mid y = 0 \sim \mathcal{N}(\mu_0, \Sigma)$
- $x \mid y = 1 \sim \mathcal{N}(\mu_1, \Sigma)$

In particular, y follows a Bernoulli distribution parameterized by ψ, $x \mid y = 0$ follows a Gaussian distribution parameterized by μ_0 and Σ, and $x \mid y = 1$ follows a Gaussian distribution parameterized by μ_1 and Σ. Specifically, we have the following equation according to these assumptions.

$$p(y) = \psi^y (1-\psi)^{1-y} \tag{2}$$

$$p(x \mid y = 0) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_0)^T \Sigma^{-1} (x-\mu_0)\right) \tag{3}$$

$$p(x \mid y = 1) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)^T \Sigma^{-1} (x-\mu_1)\right) \tag{4}$$

Given m sample data $\{(x^{(i)}, y^{(i)})\}_{i=1,\cdots,m}$, the log-likelihood is defined as

$$
\begin{aligned}
\ell(\psi, \mu_0, \mu_1, \Sigma) &= \log \prod_{i=1}^{m} p(x^{(i)}, y^{(i)}; \psi, \mu_0, \mu_1, \Sigma) \\
&= \log \prod_{i=1}^{m} p(x^{(i)} \mid y^{(i)}; \mu_0, \mu_1, \Sigma) p(y^{(i)}; \psi) \\
&= \sum_{i=1}^{m} \log p(x^{(i)} \mid y^{(i)}; \mu_0, \mu_1, \Sigma) + \sum_{i=1}^{m} \log p(y^{(i)}; \psi) \qquad (5)
\end{aligned}
$$

where ψ, μ_0, and σ are parameters. Substituting Eq. (2)~(4) into Eq. (5) gives us a full expression of $\ell(\psi, \mu_0, \mu_1, \Sigma)$

$$
\begin{aligned}
& \ell(\psi, \mu_0, \mu_1, \Sigma) \\
= & \sum_{i=1}^{m} \log p(x^{(i)} \mid y^{(i)}; \mu_0, \mu_1, \Sigma) + \sum_{i=1}^{m} \log p(y^{(i)}; \psi) \\
= & \sum_{i:y^{(i)}=0} \log \left[\frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp \left(-\frac{1}{2} (x - \mu_0)^T \Sigma^{-1} (x - \mu_0) \right) \right] \\
& + \sum_{i:y^{(i)}=1} \log \left[\frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp \left(-\frac{1}{2} (x - \mu_1)^T \Sigma^{-1} (x - \mu_1) \right) \right] \\
& + \sum_{i=1}^{m} \log \psi^{y^{(i)}} (1 - \psi)^{y^{(i)}}
\end{aligned}
$$

We then maximize the log-likelihood function $\ell(\psi, \mu_0, \mu_1, \Sigma)$ so as to get the optimal values for ψ, μ_0, and σ, such that the resulting GDA model can best fit the given training data. In particular, we let

$$
\begin{aligned}
\nabla_{\mu_0} \ell(\psi, \mu_0, \mu_1, \Sigma) = 0 \\
\nabla_{\mu_1} \ell(\psi, \mu_0, \mu_1, \Sigma) = 0 \\
\nabla_{\Sigma} \ell(\psi, \mu_0, \mu_1, \Sigma) = 1
\end{aligned}
$$

A careful derivative gives us

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

Now we can use the above results to calculate the expression of $p(y)$, $p(x \mid y = 0$, and $p(x \mid y = 1)$, according to our assumptions (2)~(4).

2 Gaussian Discriminant Analysis and Logistic Regression

By far, we introduce two classification algorithms, *Logistic Regression* (LR) and GDA. We now dive into investigation the relationship between them. Given a test data sample x, we can calculate $p(y = 1 \mid x)$ as follows

$$\begin{aligned} p(y=1 \mid x) &= \frac{p(x \mid y=1)p(y=1)}{p(x)} \\ &= \frac{p(x \mid y=1)p(y=1)}{p(x \mid y=1)p(y=1) + p(x \mid y=0)p(y=0)} \\ &= \frac{1}{1 + \frac{p(x|y=0)p(y=0)}{p(x|y=1)p(y=1)}} \end{aligned} \tag{6}$$

According to our assumptions (2)∼(4), we have

$$\begin{aligned} & \frac{p(x \mid y=0)p(y=0)}{p(x \mid y=1)p(y=1)} \\ = \ & \exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0) + \frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right) \cdot \frac{1-\psi}{\psi} \\ = \ & \exp\left((\mu_0-\mu_1)^T\Sigma^{-1}x + \frac{1}{2}\left(\mu_1^T\Sigma^{-1}\mu_1 - \mu_0^T\Sigma^{-1}\mu_0\right)\right) \cdot \exp\left(\log\left(\frac{1-\psi}{\psi}\right)\right) \\ = \ & \exp\left((\mu_0-\mu_1)^T\Sigma^{-1}x + \frac{1}{2}\left(\mu_1^T\Sigma^{-1}\mu_1 - \mu_0^T\Sigma^{-1}\mu_0\right)\right) \cdot \exp\left(\log\left(\frac{1-\psi}{\psi}\right)\right) \\ = \ & \exp\left((\mu_0-\mu_1)^T\Sigma^{-1}x + \frac{1}{2}\left(\mu_1^T\Sigma^{-1}\mu_1 - \mu_0^T\Sigma^{-1}\mu_0\right) + \log\left(\frac{1-\psi}{\psi}\right)\right) \end{aligned}$$

If we assume

$$\begin{aligned} x &:= \begin{bmatrix} x \\ 1 \end{bmatrix} \\ \theta &= \begin{bmatrix} (\mu_0-\mu_1)^T\Sigma^{-1} \\ \frac{1}{2}\left(\mu_1^T\Sigma^{-1}\mu_1 - \mu_0^T\Sigma^{-1}\mu_0\right) + \log\left(\frac{1-\psi}{\psi}\right) \end{bmatrix} \end{aligned}$$

we have

$$\begin{aligned} & \frac{p(x \mid y=0)p(y=0)}{p(x \mid y=1)p(y=1)} \\ = \ & \exp\left((\mu_0-\mu_1)^T\Sigma^{-1}x + \frac{1}{2}\left(\mu_1^T\Sigma^{-1}\mu_1 - \mu_0^T\Sigma^{-1}\mu_0\right) + \log\left(\frac{1-\psi}{\psi}\right)\right) \\ = \ & \exp\left(\theta^T x\right) \end{aligned} \tag{7}$$

By substituting (7) into Eq. (6), we finally represent $p(y = 1 \mid x)$ as

$$p(y=1 \mid x) = \frac{1}{1+\exp(\theta^T x)} \tag{8}$$

Similarly, we have

$$\begin{aligned}
& p(y=0 \mid x) \\
= \; & \frac{p(x \mid y=0)p(y=0)}{p(x)} \\
= \; & \frac{p(x \mid y=0)p(y=0)}{p(x \mid y=1)p(y=1)+p(x \mid y=0)p(y=0)} \\
= \; & \frac{1}{1+\frac{p(x|y=1)p(y=1)}{p(x|y=0)p(y=0)}} \\
= \; & \frac{1}{1+\exp\left((\mu_1-\mu_0)^T\Sigma^{-1}x+\frac{\mu_0^T\Sigma^{-1}\mu_0-\mu_1^T\Sigma^{-1}\mu_1}{2}+\log\left(\frac{\psi}{1-\psi}\right)\right)}
\end{aligned}$$

Therefore, we conclude that GDA model can be reformulated as logistic regression. But the question is, which one is better? GDA makes stronger modeling assumptions, and is more data efficient (i.e., requires less training data to learn "well") when the modeling assumptions are correct or at least approximately correct, while LR makes weaker assumptions, and is significantly more robust deviations from modeling assumptions. Hence, when the data is indeed non-Gaussian, then in the limit of large datasets, logistic regression will almost always do better than GDA. In practice, logistic regression is used more often than GDA

3 Naive Bayes

Again, we assume that the m training data are denoted by $\{x^{(i)}, y^{(i)}\}_{i=1,\cdots,m}$, where $x^{(i)}$ is a n-dimensional vector with each component $x_j^{(i)} \in \{0,1\}$ ($j = 1,\cdots,n$), and $y^{(i)} \in \{1,\cdots,k\}$. For brevity, we use $[k]$ to denote set $\{1,2,\cdots k\}$. Therefore, we have $i \in [m]$, $j \in [n]$ and $y \in [k]$. In Naive Bayes (NB) model, features and labels can be represented by random variables $\{X_j\}_{j\in[n]}$ and Y, respectively. Furthermore, for $\forall j \neq j'$, Naive Bayes assumes X_j and $X_{j'}$ are conditionally independent given Y. Therefore, we have

$$\begin{aligned}
& p(X_1=x_1, X_2=x_2, \cdots, X_n=x_n \mid Y=y) \\
= \; & \prod_{j=1}^{n} p(X_j=x_j \mid X_1=x_1, X_2=x_2, \cdots, X_{j-1}=x_{j-1}, Y=y) \\
= \; & \prod_{j=1}^{n} p(X_j=x_j \mid Y=y)
\end{aligned}$$

Moreover, $p(Y=y, X_1=x_1, \cdots, X_n=x_n)$ can be calculated as

$$\begin{aligned}
& p(Y=y, X_1=x_1, \cdots, X_n=x_n) \\
= \; & p(X_1=x_1, \cdots, X_n=x_n \mid Y=y)p(Y=y) \\
= \; & p(Y=y)\prod_{j=1}^{n} p(X_j=x_j \mid Y=y)
\end{aligned}$$

By now, we have two set of parameters: i) $p(Y = y)$ for $\forall y \in [k]$, and ii) $p(X_j = x \mid Y = y)$ for $\forall x_j \in \{0, 1\}$ where $j \in [n]$, $\forall y \in [k]$. For brevity, we make the following assumption

$$\begin{aligned} &p(Y = y) := p(y) \\ &p(X_j = x \mid Y = y) := p_j(x \mid y) \end{aligned}$$

More specifically, $p(y)$ denotes the prior probability of $Y = y$, while $p_j(x \mid y)$ denotes the posterior probability of $X_j = x$ given $Y = y$.

Given a set of m training data $\{x^{(i)}, y^{(i)}\}_{i\in[m]}$, the log-likelihood function can be defined by

$$\begin{aligned} \ell(\Omega) &= \log \prod_{i=1}^{m} p(x^{(i)}, y^{(i)}) \\ &= \sum_{i=1}^{m} \log p(x^{(i)}, y^{(i)}) \\ &= \sum_{i=1}^{m} \log \left(p(y^{(i)}) \prod_{j=1}^{n} p_j(x_j^{(i)} \mid y^{(i)}) \right) \\ &= \sum_{i=1}^{m} \log p(y^{(i)}) + \sum_{i=1}^{m} \sum_{j=1}^{n} \log p_j(x_j^{(i)} \mid y^{(i)}) \qquad (9) \end{aligned}$$

where we use Ω to represent the set of parameters. Again, we would like to maximize the above objective function with respect to $\{p(y)\}_{y\in[k]}$ and $\{p_j(x \mid y)\}_{j\in[n],x\in\{0,1\},y\in[k]}$. Mathematically, our problem can be formulated as

$$\begin{aligned} \max \quad & \ell(\Omega) = \sum_{i=1}^{m} \log p(y^{(i)}) + \sum_{i=1}^{m} \sum_{j=1}^{n} \log p_j(x_j^{(i)} \mid y^{(i)}) \\ s.t. \quad & \sum_{y=1}^{k} p(y) = 1 \\ & \sum_{x\in\{0,1\}} p_j(x \mid y) = 1, \ \forall y \in [k], j \in [n] \\ & p(y) \geq 0, \ \forall y \in [k] \\ & p_j(x \mid y) \geq 0, \ \forall y \in [k], j \in [n], x \in \{0, 1\} \end{aligned}$$

To facilitate our derivations, we next try to explicitly represent $\ell(\Omega)$ as a function of the parameters. We suppose that

$$\begin{aligned} \mathit{count}(y) &= \sum_{i=1}^{m} \mathbf{1}(y^{(i)} = y), \ \forall y \in [k] \\ \mathit{count}_j(x \mid y) &= \sum_{i=1}^{m} \mathbf{1}(y^{(i)} = y \wedge x_j^{(i)} = x), \ \forall y \in [k], \ \forall x \in \{0, 1\} \end{aligned}$$

The m training data can be divided into k groups according to the different values of y, i.e., $\{i \in [m] \mid y^{(i)} = 1\}$, $\{i \in [m] \mid y^{(i)} = 2\}$, $\cdots$, and $\{i \in [m] \mid$

$y^{(i)} = k\}$ Then, we have

$$\begin{aligned}
\sum_{i=1}^{m} \log p(y^{(i)}) &= \sum_{i:y^{(i)}=1} \log p(y^{(i)} = 1) + \cdots + \sum_{i:y^{(i)}=k} \log p(y^{(i)} = k) \\
&= \mathit{count}(y = 1) \log p(y = 1) + \cdots + \mathit{count}(y = k) \log p(y = k) \\
&= \sum_{y=1}^{k} \mathit{count}(y) \log p(y) \qquad (10)
\end{aligned}$$

Also, for $\forall y \in [k]$ and $\forall j \in [n]$, we can further divided the data set $\{i \in [m] \mid y^{(i)} = y\}$ into two groups, i.e.,

$$\mathcal{D}_{0|y}^{(j)} = \{i \in [m] \mid x_j^{(i)} = 0 \wedge y^{(i)} = y\}$$

and

$$\mathcal{D}_{1|y}^{(j)} = \{i \in [m] \mid x_j^{(i)} = 1 \wedge y^{(i)} = y\}$$

Specifically, $D_{0|y}^{(j)}$ denotes a subset of the training data with the j-th feature being 0 and the label being y, while $\mathcal{D}_{1|y}^{(j)}$ is a subset of the training data with the j-th feature being 1 and the label being y. Accordingly, we have

$$\begin{aligned}
&\sum_{i=1}^{m} \sum_{j=1}^{n} \log p_j(x_j^{(i)} \mid y^{(i)}) \\
= &\sum_{j=1}^{n} \sum_{i:y^{(i)}=1} \log p_j(x_j^{(i)} \mid y^{(i)} = 1) + \cdots + \sum_{j=1}^{n} \sum_{i:y^{(i)}=k} \log p_j(x_j^{(i)} \mid y^{(i)} = k) \\
= &\sum_{j=1}^{n} \left(\sum_{i \in \mathcal{D}_{0|1}^{(j)}} \log p_j(x_j^{(i)} = 0 \mid y^{(i)} = 1) + \sum_{i \in \mathcal{D}_{1|1}^{(j)}} \log p_j(x_j^{(i)} = 1 \mid y^{(i)} = 1) \right) \\
&\vdots \\
&+ \sum_{j=1}^{n} \left(\sum_{i \in \mathcal{D}_{0|k}^{(j)}} \log p_j(x_j^{(i)} = 0 \mid y^{(i)} = k) + \sum_{i \in \mathcal{D}_{1|k}^{(j)}} \log p_j(x_j^{(i)} = 1 \mid y^{(i)} = k) \right) \\
= &\sum_{j=1}^{n} \sum_{y=1}^{k} \sum_{x \in \{0,1\}} \mathit{count}_j(x \mid y) \log p_j(x \mid y) \qquad (11)
\end{aligned}$$

By substituting (10) and (11) into (9), we finally formulate our problem as follows

$$
\begin{aligned}
\max \quad & \ell(\Omega) = \sum_{y=1}^{k} count(y) \log p(y) + \sum_{j=1}^{n} \sum_{y=1}^{k} \sum_{x \in \{0,1\}} count_j(x \mid y) \log p_j(x \mid y) \\
s.t. \quad & \sum_{y=1}^{k} p(y) = 1 \\
& \sum_{x \in \{0,1\}} p_j(x \mid y) = 1, \ \forall y \in [k], j \in [n] \\
& p(y) \geq 0, \ \forall y \in [k] \\
& p_j(x \mid y) \geq 0, \ \forall y \in [k], j \in [n], x \in \{0,1\}
\end{aligned}
$$

where $p_j(x \mid y)$ and $p(y)$ are variables while $count(y)$ and $count_j(x \mid y)$ are knowns. By utilizing Lagrange multipliers, we resolve the above problem and get the optimal values for all parameters The maximum-likelihood estimates for Naive Bayes model are as follows

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

and

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

As discussed above, we investigate NB model by assuming $x_j \in \{0,1\}$ for $\forall j \in [n]$. What if $x \in [u]$? Can we get the same results? Check it by yourself!

After determining the values of $p_j(x \mid y)$ and $p(y)$ for $\forall j \in [n]$, $\forall x \in \{0,1\}$, and $\forall y \in [k]$, we can predict the label for a given data sample $\hat{x} = [\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_n]^T$, by calculating

$$
\begin{aligned}
& p(Y = y \mid X_1 = \hat{x}_1, \cdots, X_n = \hat{x}_n) \\
= \quad & \frac{p(X_1 = \hat{x}_1, \cdots, X_n = \hat{x}_n \mid Y = y) p(Y = y)}{p(X_1 = \hat{x}_1, \cdots, X_n = \hat{x}_n)} \\
= \quad & \frac{p(Y = y) \prod_{j=1}^{n} p(X_j = \hat{x}_j \mid Y = y)}{p(X_1 = \hat{x}_1, \cdots, X_n = \hat{x}_n)}
\end{aligned}
$$

for each $y \in [k]$. We aim at seeking for the optimal value of y such that $p(Y = y \mid X_1 = \hat{x}_1, \cdots, X_n = \hat{x}_n)$ is maximized. Specifically, since for each $y \in [k]$, $p(Y = y \mid X = x^*)$ has an identical denominator (i.e., $p(X_1 = \hat{x}_1, \cdots, X_n = \hat{x}_n)$), the label of $\hat{x}$ is defined as

$$
\hat{y} = \arg \max_{y \in \{1, \cdots, k\}} \left(p(y) \prod_{j=1}^{n} p_j(\hat{x}_j \mid y) \right)
$$

4 Naive Bayes for Multinomial Distribution

In this model, a training sample may involves a different number of features. We assume that the i-th training sample $x^{(i)}$ has n_i features. For $\forall i \in [m]$, $x^{(i)}$

has each of its features drawn from a common sample space $[v] = \{1, 2, \cdots, v\}$ identically and independently. Therefore, for $\forall j \in [n]$, we define

$$\psi(t \mid y) = p(X_j = t \mid Y = y)$$

for $\forall t \in [v]$ and $\forall y \in [k]$, which is the posterior probability of $X_j = t$ given $Y = y$. Specifically, $\psi(t \mid y)$ is the conditional probability that the j-th feature is t given that the label is y. Also, $\psi(t \mid y)$ should respect the following conditions: i) $\psi(t \mid y) \geq 0$, and ii) $\sum_{t=1}^{v} \psi(t \mid y) = 1$. We also define

$$\psi(y) = p(Y = y)$$

for $\forall y \in [k]$. We denote by Ω the set of parameters, i.e., $\Omega = \{\psi(y), \psi(t \mid y)\}_{t \in [v], y \in [k]}$

Given a set of m training data $\{(x^{(i)}, y^{(i)})\}_{i \in [m]}$, the log-likelihood function can be defined by

$$\begin{aligned}
\ell(\Omega) &= \log \prod_{i=1}^{m} p(x^{(i)}, y^{(i)}) \\
&= \log \prod_{i=1}^{m} p(x^{(i)} \mid y^{(i)}) p(y^{(i)}) \\
&= \log \prod_{i=1}^{m} \sum_{y=1}^{k} \mathbf{1}(y^{(i)} = y) p(x^{(i)} \mid y^{(i)} = y) p(y^{(i)} = y) \\
&= \sum_{i=1}^{m} \sum_{y=1}^{k} \mathbf{1}(y^{(i)} = y) \log \Big(p(x^{(i)} \mid y^{(i)} = y) p(y^{(i)} = y) \Big) \\
&= \sum_{i=1}^{m} \sum_{y=1}^{k} \mathbf{1}(y^{(i)} = y) \log \Big(p(x^{(i)} \mid y^{(i)} = y) p(y^{(i)} = y) \Big) \\
&= \sum_{i=1}^{m} \sum_{y=1}^{k} \mathbf{1}(y^{(i)} = y) \log \left(p(y^{(i)} = y) \prod_{j=1}^{n_i} p(x_j^{(i)} \mid y^{(i)} = y) \right) \\
&= \sum_{i=1}^{m} \sum_{y=1}^{k} \mathbf{1}(y^{(i)} = y) \log \left(\psi(y) \prod_{t=1}^{v} \psi(t \mid y)^{count^{(i)}(t)} \right) \\
&= \sum_{i=1}^{m} \sum_{y=1}^{k} \mathbf{1}(y^{(i)} = y) \left(\log \psi(y) + \sum_{t=1}^{v} count^{(i)}(t) \psi(t \mid y) \right)
\end{aligned}$$

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

By now, we formulate our NB model for multinomial distribution as follows

$$\begin{aligned}
\max \quad & \ell(\Omega) = \sum_{i=1}^{m} \sum_{y=1}^{k} \mathbf{1}(y^{(i)} = y) \left(\log \psi(y) + \sum_{t=1}^{v} count^{(i)}(t) \psi(t \mid y) \right) \\
s.t. \quad & \psi(y) \geq 0, \ \forall y \in [k] \\
& \psi(t \mid y) \geq 0, \ \forall t \in [v] \ \forall y \in [k] \\
& \sum_{y=1}^{k} \psi(y) = 1, \\
& \sum_{t=1}^{v} \psi(t \mid y) = 1, \ \forall y \in [k]
\end{aligned}$$

By Lagrange multiplier, we get the optimal solution

$$\begin{aligned}
\psi(t \mid y) &= \frac{\sum_{i=1}^{m} \mathbf{1}(y^{(i)} = y) count^{(i)}(t)}{\sum_{i=1}^{m} \mathbf{1}(y^{(i)} = y) \sum_{t=1}^{v} count^{(i)}(t)} \\
\psi(y) &= \frac{\sum_{i=1}^{m} \mathbf{1}(y^{(i)} = y)}{m}
\end{aligned}$$

5 Expectation-Maximization Algorithm

We hereby look at *Expectation-Maximization* (EM) algorithm.