Logistic Model, Bayesian Model and EM

crackhopper

Tue Dec 13 20:07:43 2016

Outline

Problem

We have a set of data $D = \{(x_i, y_i)\}_{i=1}^N$, where $x_i \in \mathbb{R}^m$ and $y_i \in \{0, 1\}$

Question

- How do we train a classifier?

Assunmption

We assume a linear model, $z = w' \cdot x$.

- If $y = 0$, we assume $z \sim \mathcal{N}(\mu_0, \sigma^2)$, i.e. $p(z|y = 0) = \mathcal{N}(\mu_0, \sigma^2)$
- If $y = 1$, we assume $z \sim \mathcal{N}(\mu_1, \sigma^2)$, i.e. $p(z|y = 1) = \mathcal{N}(\mu_1, \sigma^2)$

The Logistic Model

So we have $p(z|y)$, and we can get $p(y|z)$, which is what we want, by Bayesian Theorem:

$$
\begin{aligned}
p(y=0|z) &= \frac{p(z|y=0)p(y=0)}{p(z|y=0)p(y=0)+p(z|y=1)p(y=1)} \\
&= \frac{e^{\frac{(w'\cdot x-\mu_0)^2}{2\sigma^2}}}{e^{\frac{(w'\cdot x-\mu_0)^2}{2\sigma^2}}+e^{\frac{(w'\cdot x-\mu_1)^2}{2\sigma^2}}} \\
&= \frac{1}{1+e^{\frac{(w'\cdot x-\mu_1)^2-(w'\cdot x-\mu_0)^2}{2\sigma^2}}} \\
&= \frac{1}{1+e^{C(\mu_1^2-\mu_0^2-(\mu_1-\mu_0)w'\cdot x)}} \\
&= \frac{1}{1+e^{-\mathbf{w}\cdot\mathbf{x}}}
\end{aligned}
$$

where $\mathbf{w}=(w_0,...,w_m), \mathbf{x}=(1,x_1,...,x_m)$

Optimization Target

The function

$$p(y=0|z) = \sigma(z) = \frac{1}{1+e^{-z}}$$

is called the sigmoid function, which transform a set of variables into a probability between $(0, 1)$.

So, we try to maximize likelihood. The optimization target for logistic regression is

$$\begin{aligned} & \max_{w,b} \prod [\sigma(\mathbf{w} \cdot \mathbf{x}_i)]^{y_i} [1 - \sigma(\mathbf{w} \cdot \mathbf{x}_i)]^{1-y_i} \\ = \; & \max_{w,b} \sum [y_i \log(\sigma_i) + (1 - y_i) \log(1 - \sigma_i)] \end{aligned}$$

Solve The Target

Remember the Newton-Raphson Method: $x_{k+1} = x_k - \frac{f(x_k)}{f'_{x_k}}$, it is same here, let the partial derivative of the log-loss function be zero, and we have:

$$\begin{aligned} & \sum_i \frac{\partial}{\partial \sigma_i}[y_i \log(\sigma_i) + (1 - y_i) \log(1 - \sigma_i)] \\ = & \sum_i y_i \frac{1}{\sigma_i} \sigma_i (1 - \sigma_i) - (1 - y_i) \frac{1}{1 - \sigma_i} \sigma_i (1 - \sigma_i) \\ = & \sum_i y_i (1 - \sigma_i) - (1 - y_i) \sigma_i \\ = & \sum_i y_i - \sigma_i \end{aligned}$$

and it's easy to find: $\frac{\partial \sigma_i}{\partial \mathbf{w}} = (1, x_{i1}, ..., x_{im})^T$

Solve The Target

so we can write the condition as

$$\begin{bmatrix} 1 & 1 & \dots & 1 \\ x_{11} & x_{21} & \dots & x_{n1} \\ \vdots & \vdots & \dots & \vdots \\ x_{11} & x_{21} & \dots & x_{n1} \end{bmatrix} \begin{bmatrix} y_1 - \sigma_1 \\ y_2 - \sigma_2 \\ \vdots \\ y_n - \sigma_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Define the left part to be $F(\mathbf{w}) = X^T(\mathbf{y} - \sigma)$, we further need to get the derivative of F, i.e. the second order derivative of

$$L(w) = \sum [y_i \log(\sigma_i) + (1 - y_i) \log(1 - \sigma_i)]$$

Solve The Target

It can be calculated as

$$\begin{aligned} H(\mathbf{w}) &= \frac{\partial}{\partial \sigma} \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_{11} & x_{21} & \dots & x_{n1} \\ \vdots & \vdots & \dots & \vdots \\ x_{11} & x_{21} & \dots & x_{n1} \end{bmatrix} \begin{bmatrix} y_1 - \sigma_1 \\ y_2 - \sigma_2 \\ \vdots \\ y_n - \sigma_n \end{bmatrix} \frac{\partial \sigma}{\partial \mathbf{w}} \\ &= X^T V X \end{aligned}$$

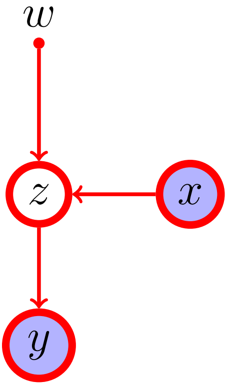
where $V = \text{diag}(-\sigma_i(1 - \sigma_i))$

so use Newton method to find the zero point,

$$\mathbf{w}_{new} = \mathbf{w}_{old} + (H(w))^{-1} F(\mathbf{w})$$

Logistic Model, Bayesian Model and EM

Logistic Model (revisit)



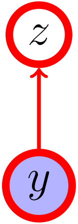
We use graphic notation:

- circle : random variables
- shaded : observed random variables
- point : parameters
- line : relationship

Example: The left graph means, we can have the joint distribution of z and y by

$$p(x, y, z|w) = p(y|z)p(z|x, w)p(x)$$

It is the graph of our logistic model.



We assume

- likelihood $p(z|y)$ is a Gaussian distribution
- prior $p(y)$ is equally distributed.

then we try to calculate

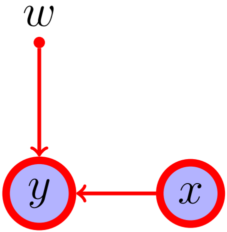
- posterior $p(y|z)$

using

$$p(y=0|z) = \frac{p(z|y=0)p(y=0)}{\sum_k p(z|y=k)p(y=k)} = \sigma(z)$$

So we get the logistic function by a equally distributed prior $p(y)$, and a pre-defined likelihood $p(z|y)$.

Maximize Likelihood

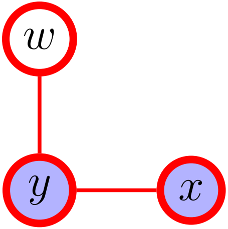


Now is the graph when we solve the logistic model. We want to obtain the parameter w. So we do a Maximize Marginal Likelihood.

The likelihood function

$$
\begin{aligned}
& p(Y|X,w) \\
= & \prod_{n=1}^{N} p(y_n|x_n,w) \\
= & \prod_{n=1}^{N} (\sigma(w \cdot x_n))^{y_n} (1-\sigma(w \cdot x_n))^{1-y_n}
\end{aligned}
$$

Logistic Model (modified)



Now we treat w as a random variable.

- the likelihood is $p(Y, X|w) = p(Y|X, w)p(X)$
- the prior is be $p(w)$

Maximize A Posterior

Another common method is to calculate the posterior $p(w|X,Y)$, and try to maximize it, this method is called Maximize A Posterior(MAP).

By Bayesian Theorem,

$$p(w|X,Y) = \frac{p(X,Y|w)p(w)}{p(X,Y)} = \frac{p(Y|X,w)p(w)}{p(Y|X)}$$

so to maximize it is equivalent to

$$\begin{aligned} & \max_{w,b} \prod [\sigma(\mathbf{w}\cdot\mathbf{x}_i)]^{y_i}[1-\sigma(\mathbf{w}\cdot\mathbf{x}_i)]^{1-y_i} p(w) \\ = & \max_{w,b} \sum [y_i \log(\sigma_i) + (1-y_i)\log(1-\sigma_i) + \log(p(w))] \end{aligned}$$

The only difference is we treat the parameter as a random variable which has a prior distribution. (It finally become a regularizer in the optimization target)

Problem Redefine

We have a set of data $D = \{(\mathbf{x}_i)\}_{i=1}^{N}$. We don't have any label. But we assume there is some label $\mathbf{y}_i$ that we cannot observed. The unobserved random variable is called the latent variable.

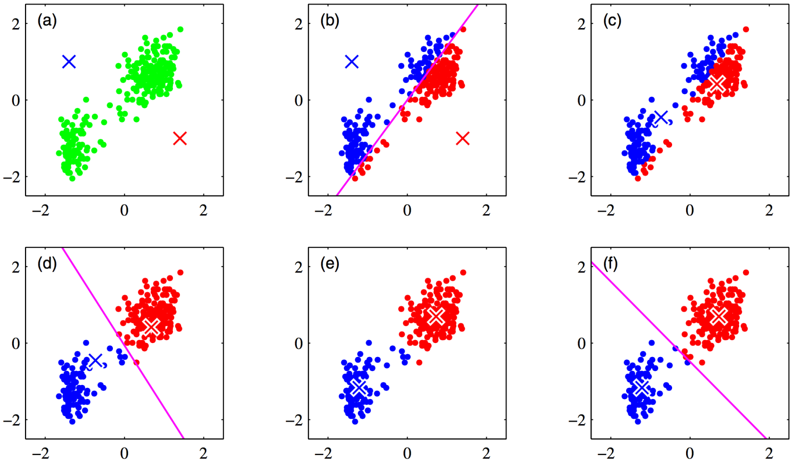
K-Means Clustering

We introduce a corresponding set of binary indicator variables $r_{nk} \in \{0, 1\}$, so that if x_n is assigned to cluster k, $r_{nk} = 1$, otherwise $r_{nk} = 0$. We then define an objective function, (also called distortion measure)

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||x_n = \mu_k||^2$$

1. keeping μ_k fixed, we minimize J with respect to r_{nk}
2. keeping r_{nk} fixed, we minimize J with respect to μ_k

K-Means: Illustration



Mixtures of Gaussians

Let $\mathbf{y}$ to be one-hot vector of K class label. (i.e. $y_k \in \{0, 1\}$ and $\sum_k y_k = 1$), and we denote $p(y_k = 1) = \pi_k$. so the distribution of $\mathbf{y}$ can be written as

$$p(\mathbf{y}) = \prod_{k=1}^{K} \pi_k^{y_k}$$

Similarly, the conditional distribution of $\mathbf{z}$ given a particular value for $\mathbf{y}$ is a Gaussian

$$p(\mathbf{z}|y_k = 1) = \mathcal{N}(\mathbf{z}|\mu_k, \mathbf{\Sigma}_k)$$

which can also be written

$$p(\mathbf{z}|\mathbf{y}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{z}|\mu_k, \mathbf{\Sigma}_k)^{y_k}$$

Mixtures of Gaussians

The marginal distribution of $\mathbf{z}$ is

$$p(\mathbf{z}) = \sum_{\mathbf{y}} p(\mathbf{y})p(\mathbf{z}|\mathbf{y}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{z}|\mu_k, \mathbf{\Sigma}_k)$$

which is a mixture of Gaussian.

Maximum likelihood

Suppose we have a data set of observations $\mathbf{z}_1, \mathbf{z}_N$, we can denote the data as $N \times D$ matrix $\mathbf{X}$. So the likelihood would be

$$\ln p(\mathbf{Z}|\pi, \mu, \mathbf{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{z}_n|\mu_k, \Sigma_k) \right\}$$

EM for Gaussian mixtures

Recall the Maximum Likelihood

$$\ln p(\mathbf{Z}|\pi, \mu, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{z}_n | \mu_k, \Sigma_k) \right\}$$

Setting the derivatives of above equation with respect to the means μ_k to zero,

$$0 = -\sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(\mathbf{z}_n | \mu_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{z}_n | \mu_j, \boldsymbol{\Sigma}_j)} \boldsymbol{\Sigma}_k (\mathbf{z}_n - \mu_k)$$

And we denote the posterior $\gamma(y_{nk})$

$$\gamma(y_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{z}_n | \mu_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{z}_n | \mu_j, \boldsymbol{\Sigma}_j)}$$

Firstly, we can calculate $\gamma(y_{nk})$ from data, then we just fix it to maximize the other parameters.

EM for Gaussian mixtures

After we substitute the posterior, we got the equation to be,

$$0 = -\sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{\Sigma}_k (\mathbf{z}_n - \mu_k)$$

(Note, the $\gamma(\mathsf{z}_{\mathsf{nk}})$ is calculated by using the old $\mu_k, \boldsymbol{\Sigma}_k$). Rearrange above we obtain the new μ_k

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(y_{nk}) \mathbf{z}_n$$

where $N_k = \sum_{n=1}^{N} \gamma(y_{nk})$
Samiliar, for the derivative of covariance matrix, we have

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(y_{nk}) (\mathbf{z}_n - \mu_k)(\mathbf{z}_n - \mu_k)^T$$

EM for Gaussian mixtures

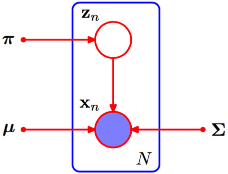
Finally, we maximize $\ln p(\mathbf{Z}|\pi, \mu, \mathbf{\Sigma})$ with respect to the mixing coefficients π_k, with the constraint $\sum_k \pi_k = 1$, we have the optimization target:

$$\ln p(\mathbf{Z}|\pi, \mu, \mathbf{\Sigma}) + \lambda(\sum_{k=1}^{K} \pi_k - 1)$$

which gives

$$0 = \sum_{n=1}^{N} \frac{\mathcal{N}(\mathbf{z}_n|\mu_k, \mathbf{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{z}_n|\mu_j), \mathbf{\Sigma}_j)} + \lambda$$

which we find $\lambda = -N$ and $\pi_k = \frac{N_k}{N}$



Expectation Step (E Step)

we use the current values for the parameters to evaluate the posterior probabilities, or responsibilitise by

$$\gamma(y_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{z}_n | \mu_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{z}_n | \mu_j, \boldsymbol{\Sigma}_j)}$$

(calculate the expectation of y from data)

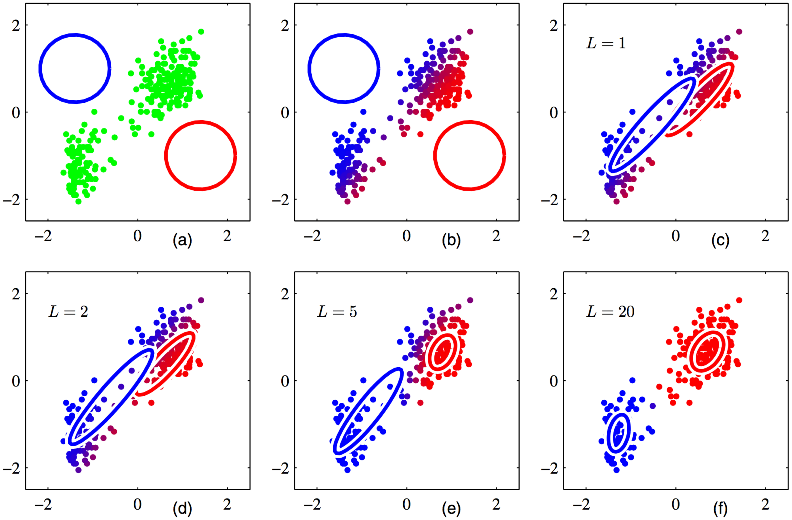
EM for Gaussian mixtures : Summary

Maximization Step (M Step)

Maximize the likelihood to re-estimate the means, covariance, and mixing coefficients,

$$
\begin{aligned}
\mu_k &= \frac{1}{N_k} \sum_{n=1}^{N} \gamma(y_{nk}) \mathbf{z}_n \\
\Sigma_k &= \frac{1}{N_k} \sum_{n=1}^{N} \gamma(y_{nk}) (\mathbf{z}_n - \mu_k)(\mathbf{z}_n - \mu_k)^T \\
\pi_k &= \frac{N_k}{N}
\end{aligned}
$$

EM : Illustration



An Alternative View of EM

We now use $\mathbf{X}$ to denote the data and $\mathbf{Z}$ to denote the latent variables. The goal of the EM algorithm is to find maximum likelihood solutions for models having latent variables.

The log-likelihood:

$$\ln p(\mathbf{X}|\theta) = \ln \left\{ \sum_{\mathbf{z}} p(\mathbf{X}, \mathbf{Z}|\theta) \right\}$$

Note: summation is inside the log function.

The General EM Algorithm

E-Step

Calculating the posterior $p(\mathbf{Z}|\mathbf{X}, \theta_{old})$.

And we further got an expectation of the optimization target over a posterior

$$Q(\theta, \theta_{old}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta_{old}) \ln p(\mathbf{X}, \mathbf{Z}|\theta)$$

M-Step

We maximize the expectation $Q(\theta, \theta_{old})$. And because now the summation is outside of the log function, it is much easier to calculate.

$$\theta_{new} = \arg\max_{\theta} Q(\theta, \theta_{old})$$

Summary : General EM Algorithm

At first, we maximize the likelihood based on incomplete data (i.e. The marginal distribution)

$$\max_{\theta} \ln p(\mathbf{X}|\theta) = \max_{\theta} \ln \left\{ \sum_{\mathbf{z}} p(\mathbf{X}, \mathbf{Z}|\theta) \right\}$$

In General EM algorithm, we maximize the likelihood based on the complete data (i.e. the joint distribution)

$$\max_{\theta} \ln p(\mathbf{X}, \mathbf{Z}|\theta)$$

But we don't have $\mathbf{Z}$, so we calculate the posterior of $\mathbf{Z}$, and average the target over the posterior

$$\max_{\theta} \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta_{old}) \ln p(\mathbf{X}, \mathbf{Z}|\theta)$$

Gaussian mixtures revisited

We would get the same result by using General EM Algorithm.

$$
\begin{aligned}
\frac{\partial}{\partial\theta}\ln\left\{\sum_{\mathbf{z}} p(\mathbf{X},\mathbf{Z}|\theta)\right\} &= \sum_{\mathbf{z}} \frac{1}{\sum_{\mathbf{z}} p(\mathbf{X},\mathbf{Z}|\theta)} \frac{\partial}{\partial\theta} p(\mathbf{X},\mathbf{Z}|\theta) \\
&= \sum_{\mathbf{z}} \frac{p(\mathbf{X},\mathbf{Z}|\theta)}{\sum_{\mathbf{z}} p(\mathbf{X},\mathbf{Z}|\theta)} \frac{1}{p(\mathbf{X},\mathbf{Z}|\theta)} \frac{\partial}{\partial\theta} p(\mathbf{X},\mathbf{Z}|\theta)
\end{aligned}
$$

and we substitute the $\frac{p(\mathbf{X},\mathbf{Z}|\theta)}{\sum_{\mathbf{z}} p(\mathbf{X},\mathbf{Z}|\theta)}$ with $p(\mathbf{Z}|\mathbf{X},\theta_{old})$

$$
\frac{\partial}{\partial\theta}\ln\left\{\sum_{\mathbf{z}} p(\mathbf{X},\mathbf{Z}|\theta)\right\} = \sum_{\mathbf{Z}} \frac{p(\mathbf{Z}|\mathbf{X},\theta_{old})}{p(\mathbf{X},\mathbf{Z}|\theta)} \frac{\partial}{\partial\theta} p(\mathbf{X},\mathbf{Z}|\theta)
$$

Gaussian mixtures revisited

On the other hand, for the general EM algorithm, we get the derivative as follow:

$$\frac{\partial}{\partial\theta}\sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\theta_{old}) \ln p(\mathbf{X},\mathbf{Z}|\theta) = \sum_{\mathbf{Z}} \frac{p(\mathbf{Z}|\mathbf{X},\theta_{old})}{p(\mathbf{X},\mathbf{Z}|\theta)} \frac{\partial}{\partial\theta} p(\mathbf{X},\mathbf{Z}|\theta)$$

It is the same result when compared with previous page.

Relation to K-Means