Introduction to Machine Learning

Maximum Likelihood and Bayesian Inference

Lecturers: Eran Halperin, Yishay Mansour, Lior Wolf

2013-14

- We know that X ~ B(n,p), but we do not know p.
- We get a random sample from X, a random number m.

$$\Pr(X = m \mid X \sim B(n,p)) = \binom{n}{m} p^m (1-p)^{n-m}$$

- We know that X ~ B(n,p), but we do not know p.
- We get a random sample from X, a random number m.
- The likelihood is defined as:

$$L(p; X = m) = \Pr(X = m \mid X \sim B(n, p))$$

The Likelihood Function

- Assume we have a set of hypotheses to choose from.
- Normally a hypothesis will be defined by a set of parameters θ.
- We do not know θ, but we make some observations and get data D.
- The likelihood of θ is L(θ;D) = Prob(D| θ). We are interested in the hypothesis that maximizes the likelihood.

Example

- We know that X ~ B(n,p), but we do not know p. We get a random sample from X, a random number m.
- In this case, the data D is the number m, and the parameter θ is p.
- The likelihood is

$$L(p;X=m)=\Pr(X=m\mid X\sim B(n,p))=\binom{n}{m}p^{m}(1-p)^{n-m}$$

Maximum Likelihood Estimate

$$Maximum\,likelihood = \arg\max_{\theta} L(\theta;D)$$

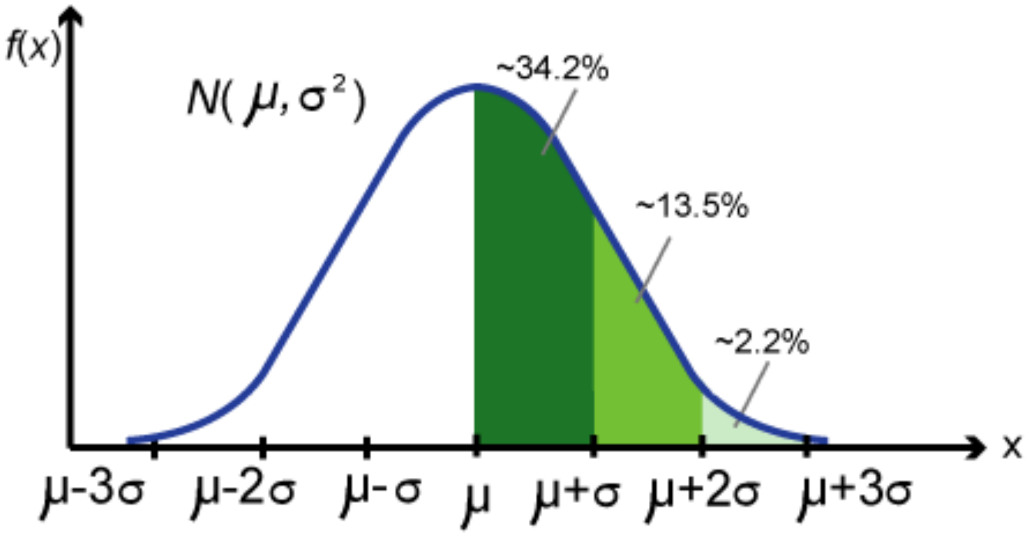
In the example above, the maximum is obtained

for $\hat{p} = \dfrac{m}{n}$

Reminder: The Normal Distribution

$$X \sim N(\mu, \sigma^2)$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Reminder: The Normal Distribution

We obtain a set of n independent samples:

$$x_1, \ldots, x_n \sim N(\mu, \sigma^2)$$

We want to estimate the model parameters: μ, σ.

Reminder: The Normal Distribution

We obtain a set of n independent samples:

$$x_1, \ldots, x_n \sim N(\mu, \sigma^2)$$

We want to estimate the model parameters: μ, σ.

$$L(\mu, \sigma; x_1, \ldots, x_n) = Pr(x_1, \ldots, x_n | \mu, \sigma) =$$

$$= \prod_{i=1}^{n} f(x_i) = \frac{1}{\sigma^n (2\pi)^{n/2}} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2}$$

Maximum Likelihood Estimate (MLE)

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \hat{\mu})^2}{n}}$$

Example

$$X_1,\ldots,X_n \sim U(0,\theta)$$

What is the maximum likelihood?

Example

$X_1,\ldots,X_n \sim U(0,\theta)$

What is the maximum likelihood?

Assume $X_{(1)} < \ldots < X_{(n)}$

For $\theta < X_{(n)}$, $\quad L(\theta;D) = 0$

For $\theta \geq X_{(n)}$, $\quad L(\theta;D) = \dfrac{1}{\theta^n}$

Max Likelihood : $\hat{\theta} = X_{(n)}$

Example: MLE of a Multinomial

- We are given a universe of possible strings (e.g., words of a language): $h_1, \ldots, h_t \in \{0,1\}^k$
- Assume a model by which the strings are generated from a multinomial with (unknown) probabilities
  $p_1, \ldots, p_t$
- We are given a sample from the multinomial with counts $c_1, \ldots, c_t$

Generative Model

$p_1 = 1/4$	01000010	
$p_2 = 1/2$	11111111	
$p_3 = 1/8$	00001111	
$p_4 = 1/8$	00000000	

11111111
00001111
01000010
11111111
00000000
11111111
01000010

Unknown

GOAL

Generative Model

$p_1 = 1/4$	01000010
$p_2 = 1/2$	11111111
$p_3 = 1/8$	00001111
$p_4 = 1/8$	00000000

11111111
00001111
01000010
11111111
00000000
11111111
01000010

$c_1 = 2$
$c_2 = 3$
$c_3 = 1$
$c_4 = 1$

Unknown

GOAL

MLE of a Multinomial

- Strings: $h_1, \ldots, h_t \in \{0,1\}^k$
- Counts: $c_1, \ldots, c_t$

$$L(p_1,\ldots,p_t; c_1,\ldots,c_t) = \binom{n}{c_1}\binom{n-c_1}{c_2}\cdots\binom{n-c_1-\ldots-c_{t-1}}{c_t} p_1^{c_1} p_2^{c_{21}} \cdots p_t^{c_t}$$

$$Max \quad \sum_i c_i \log(p_i)$$

$$s.t \sum p_i = 1, p_i > 0$$

Using Lagrange Multipliers

We are interested in maximizing:

$$Max \quad \sum_i c_i \log(p_i)$$

$$s.t \sum p_i = 1, p_i > 0$$

Instead, we will consider the Lagrange function:

$$\max \sum_i c_i \log(p_i) + \lambda(1 - \sum_i p_i), \;\; s.t. \; p_i > 0$$

An optimal solution of the original problem corresponds to a stationary point of the Lagrange function.

Using Lagrange Multipliers

$$f(\vec{p}, \lambda) = \sum_i c_i \log(p_i) + \lambda(1 - \sum_i p_i)$$

Compute the gradient:

$$\frac{\partial f}{\partial p_i} = \frac{c_i}{p_i} - \lambda \qquad \frac{\partial f}{\partial \lambda} = 1 - \sum_i p_i$$

Equating to zero:

$$p_i = \frac{c_i}{\lambda}, \lambda = \sum_i c_i = n$$

Bayesian Estimators

- Maximum likelihood: $\max Pr(D \mid \theta)$
- Advantage: No assumptions made on the model distribution.
- Disadvantage: In reality we are looking for:

$$\max Pr(\theta \mid D)$$

Is it well defined?

Prior and Posterior

Sometimes we know something about the *PRIOR* distribution $Pr(\theta)$

Then, based on Bayes rule, we can calculate the *POSTERIOR* distribution:

$$Pr(\theta \mid D) = \frac{Pr(D \mid \theta) Pr(\theta)}{Pr(D)}$$

MAP (Maximum a posteriori)

Maximum a posteriori estimation (MAP) is the mode of the posterior distribution:

$$\hat{\theta}_{MAP} = \arg\max Pr(\theta \mid D)$$

$$\hat{\theta}_{ML} = \arg\max Pr(D \mid \theta)$$

MAP (Maximum a posteriori)

Maximum a posteriori estimation (MAP) is the mode of the posterior distribution:

$$\hat{\theta}_{MAP} = \arg\max Pr(\theta \mid D)$$

$$\hat{\theta}_{ML} = \arg\max Pr(D \mid \theta)$$

$$\hat{\theta}_{MAP} = \arg\max Pr(D \mid \theta) Pr(\theta)$$

Example

Assume: $x_1, \ldots, x_n \sim N(\mu, 1)$

$$\hat{\mu}_{ML} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Normal Prior

Assume prior $\mu \sim N(0,1)$

$$\log(Pr(x_1, \ldots, x_n \mid \mu)) = -\frac{n}{2}\log(2\pi) - \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2}$$

$$\log(Pr(\mu)) = -\frac{1}{2}\log(2\pi) - \frac{\mu^2}{2}$$

$$\hat{\mu}_{MAP} = \arg\max_{\mu}\{-\mu^2 - \sum_{i=1}^{n}(x_i - \mu)^2\}$$

Normal Prior

Assume prior $\mu \sim N(0,1)$

$$\hat{\mu}_{MAP} = \arg\max_{\mu} \{-\mu^2 - \sum_{i=1}^{n}(x_i - \mu)^2\}$$

$$\hat{\mu}_{MAP} = \frac{\sum_{i=1}^{n} x_i}{n+1} \qquad \hat{\mu}_{ML} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Posterior of a Normal Prior

Assume prior $\mu \sim N(0,1)$

$$Pr(\mu \mid x_1, \ldots, x_n) \propto exp\left(-\frac{\left(\mu - \sum_{i=1}^{n} \frac{x_i}{n+1}\right)^2}{\frac{2}{n+1}}\right)$$

$$\mu \sim N\left(\frac{\sum_{i=1}^{n} x_i}{n+1}, \frac{1}{n+1}\right)$$

Choosing a prior for B(n,p)

$X \sim B(n, p)$

One sample: $X = m$

$$\hat{p}_{ML} = \frac{m}{n}$$

The Beta Distribution

$$X \sim Beta(\alpha, \beta) \qquad \alpha > 0, \beta > 0$$

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$$

$$\mu = E[X] = \frac{\alpha}{\alpha + \beta}$$

Posterior with a Beta Prior

$$X \sim B(n, p)$$

$$Assume\ prior:\ \ p \sim Beta(\alpha, \beta)$$

$$Pr(p \mid X = m, \alpha, \beta) \propto \binom{n}{m} p^m (1-p)^{n-m} \cdot \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha, \beta)}$$

$$\Downarrow$$

$$Pr(p \mid X = m, \alpha, \beta) \propto p^{m+\alpha-1}(1-p)^{n-m+\beta-1}$$

Posterior with a Beta Prior

$$Pr(p \mid X = m, \alpha, \beta) \propto p^{m+\alpha-1}(1-p)^{n-m+\beta-1}$$

$$Pr(p \mid X = m, \alpha, \beta) \sim Beta(m+\alpha, n-m+\beta)$$

$$\hat{p}_{MAP} = \frac{m+\alpha-1}{n+\alpha+\beta-2}$$

If the prior distribution is Beta then the posterior distribution is Beta as well.

A conjugate prior.

Classification (Naïve Bayes)

Cholesterol level	Heart Attack (HA)
x_1	1
x_2	1
x_3	0
x_4	1
x_5	0
x_6	0
x_7	0

Given a new individual, can we predict whether the individual will get a heart attack Based on his cholesterol level?

Classification (Naïve Bayes)

Cholesterol level	Heart Attack (HA)
x_1	1
x_2	1
x_3	0
x_4	1
x_5	0
x_6	0
x_7	0

Given a new individual, can we predict whether the individual will get a heart attack Based on his cholesterol level?

Assumption: Cholesterol levels are normally distributed with a different mean in the 1 and 0 sets.

$$Pr(x \mid HA = 1) \sim N(\mu_1, \sigma^2)$$

$$Pr(x \mid HA = 0) \sim N(\mu_0, \sigma^2)$$

μ_0, μ_1, σ can be estimated using MLE

Classification (Naïve Bayes)

$$\frac{Pr(HA = 1 \mid x)}{Pr(HA = 0 \mid x)} = \frac{Pr(HA = 1)}{Pr(HA = 0)} e^{\frac{(x-\mu_0)^2 - (x-\mu_1)^2}{2\sigma^2}}$$

Decision rule:

$$\log\left(\frac{Pr(HA = 1)}{Pr(HA = 0)}\right) + \frac{(x-\mu_0)^2 - (x-\mu_1)^2}{2\sigma^2} > 0$$

Multiple Variables

x_1	x_2	…	x_n	y
195	17	…	117	1
195	24	…	114	1
184	13	…	117	0
250	22	…	111	1
173	15	…	108	0
185	18	…	145	0
178	22	…	136	0

Assumptions:

1. Normal marginal distributions
2. Variables are independent

$$Pr(x_i \mid y = k) \sim N(\mu_{ik}, \sigma_i^2)$$

Multiple Variables

$$
\begin{aligned}
Pr(y=1 \mid x_1,\ldots,x_n) &= \frac{Pr(x_1,\ldots,x_n \mid y=1)Pr(y=1)}{Pr(x_1,\ldots,x_n)} \\
&= \frac{Pr(x_1 \mid y=1)\cdots Pr(x_n \mid y=1)Pr(y=1)}{Pr(x_1,\ldots,x_n)}
\end{aligned}
$$

Multiple Variables

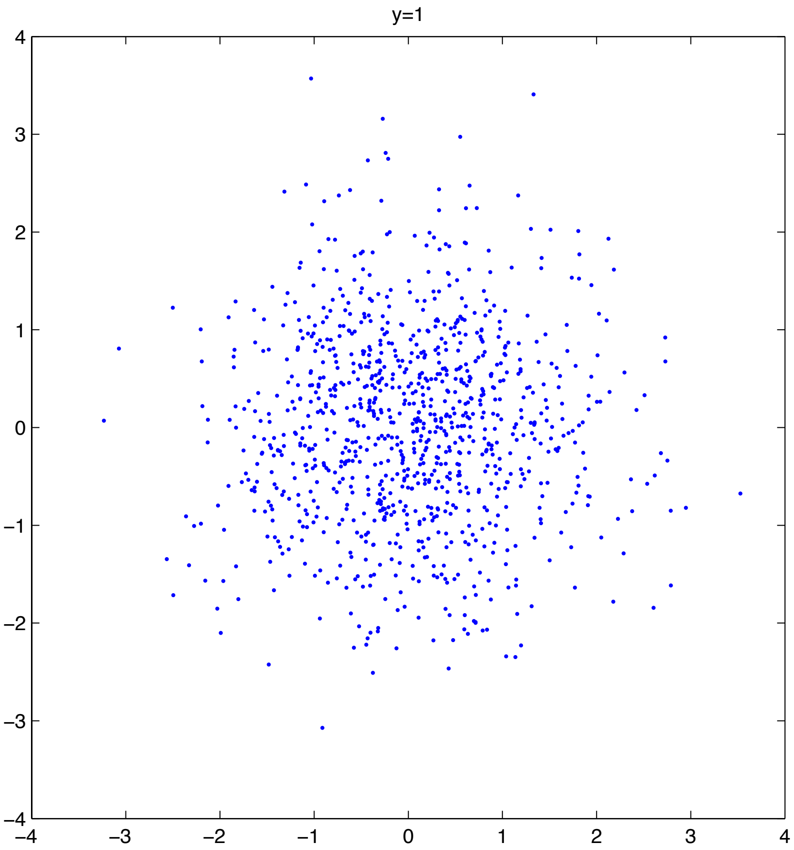
$$
\begin{aligned}
Pr(y=1 \mid x_1,\ldots,x_n) &= \frac{Pr(x_1,\ldots,x_n \mid y=1)Pr(y=1)}{Pr(x_1,\ldots,x_n)} \\
&= \frac{Pr(x_1 \mid y=1)\cdots Pr(x_n \mid y=1)Pr(y=1)}{Pr(x_1,\ldots,x_n)}
\end{aligned}
$$

$$
\begin{aligned}
\log\frac{Pr(y=1 \mid x_1,\ldots,x_n)}{Pr(y=0 \mid x_1,\ldots,x_n)} &= \log\frac{Pr(y=1)}{Pr(y=0)} + \sum_i \log\frac{Pr(x_i \mid y=1)}{Pr(x_i \mid y=0)} \\
&= \log\frac{Pr(y=1)}{Pr(y=0)} + \sum_i \frac{(x-\mu_{i0})^2-(x-\mu_{i1})^2}{2\sigma_i^2}
\end{aligned}
$$

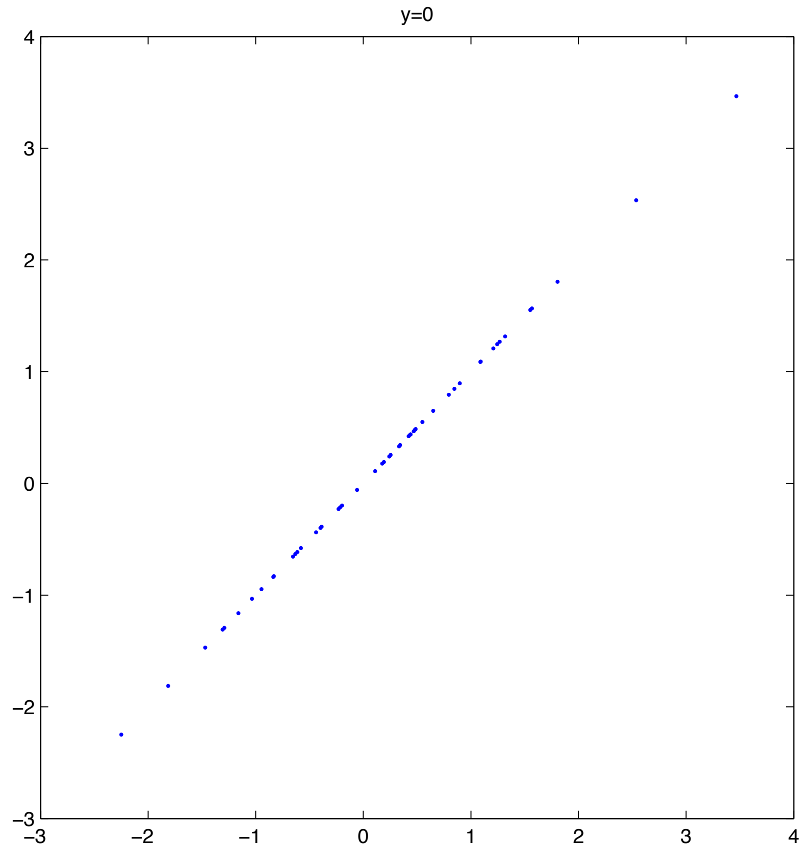
Naïve Bayes

- A Naïve assumption.
- Easy to implement.
- Often works in practice.
- Interpretation: A weighted sum of evidence.
- Allows for the incorporation of features of different distributions.
- Requires small amounts of data

Naïve Bayes Might Break...



y=1: Independent variables

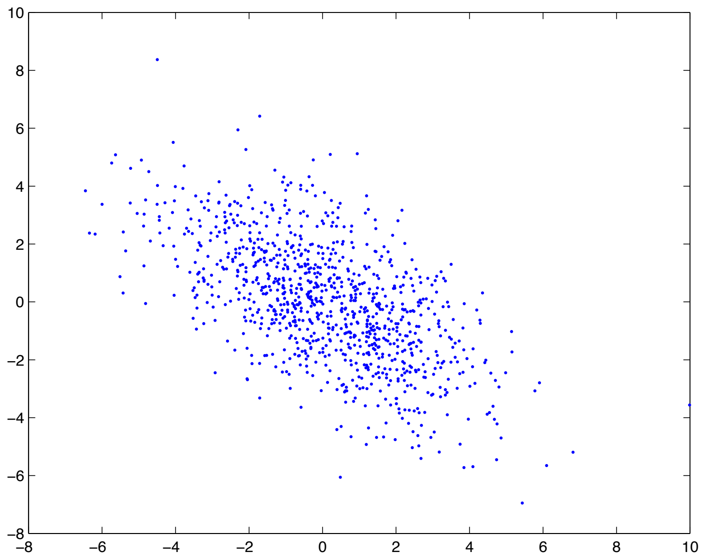


y=0: x_2=x_1

The Multivariate Normal Distribution

$$z_1, \ldots, z_n \sim N(0,1)$$

$x = Az + \mu$ is a multivariate normal distribution



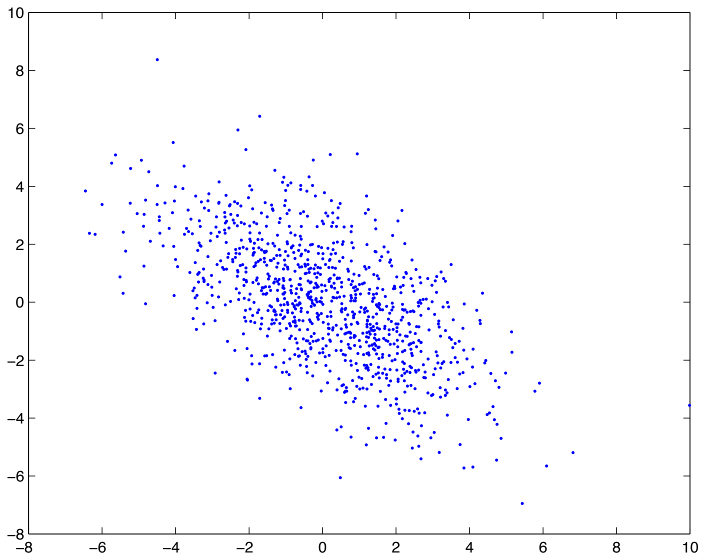
The Multivariate Normal Distribution

$$z_1, \ldots, z_n \sim N(0,1)$$

$x = Az + \mu$ is a multivariate normal distribution

Example: $A = \begin{pmatrix} 2 & 1 \\ -2 & 1 \end{pmatrix}$

$$\Sigma = AA^t = \begin{pmatrix} 5 & -3 \\ -3 & 5 \end{pmatrix}$$



The Multivariate Normal Distribution

- Notation: $X \sim MVN(\mu, \Sigma)$
- The variance-covariance matrix is Σ

$$f(x) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^t \Sigma^{-1}(x-\mu)}$$

- If we do not use Naïve Bayes we need to estimate $O(k^2)$ parameters.

Reminder: K-means objective

Given:

- Vectors $x_1, \ldots, x_n$
- A number K

Objective:

$$\min_{\mu_1\ldots,\mu_K,S_1,\ldots,S_K} \sum_{i=1}^{n} \sum_{j\in S_i} \|x_j - \mu_i\|^2$$

K-Means: A Likelihood Formulation

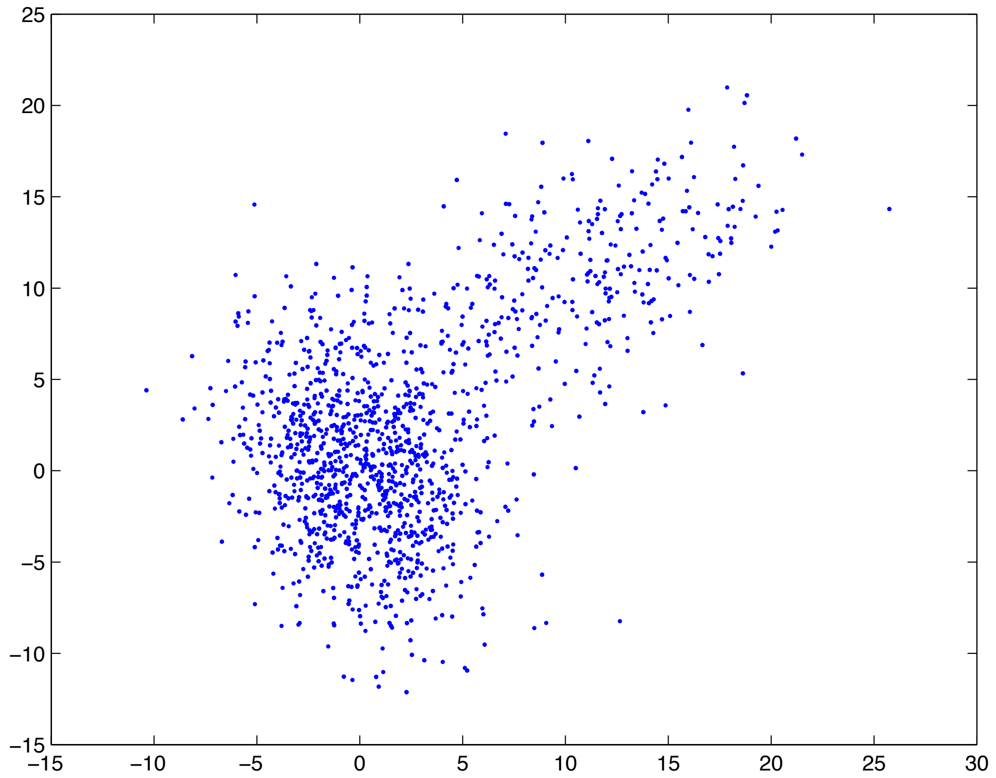
- There are unknown clusters: $S_1, \ldots, S_k$.
- The points in S_i are distributed $MVN(\mu_i, I)$
- Each point x_i originates from a cluster c_i.

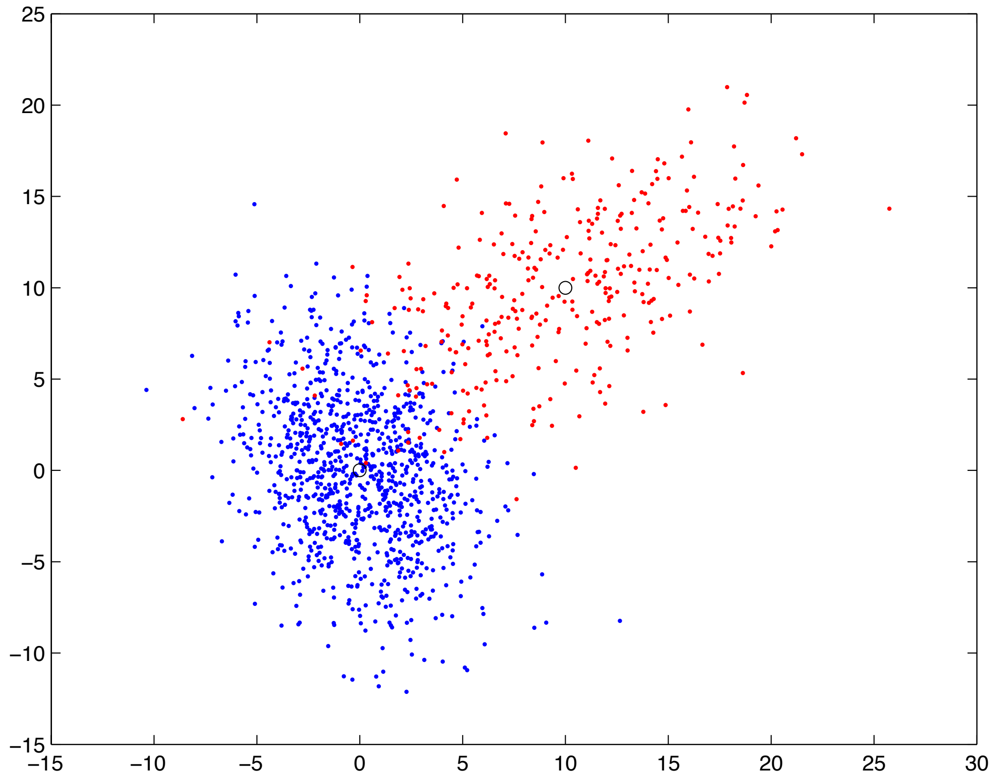
$$\theta = (c_1, \ldots, c_n, \mu_1, \ldots, \mu_k)$$

$$\log L(\theta; x_1, \ldots, x_n) = constant - \sum_{i=1}^{n} \|x_i - \mu_{c_i}\|^2$$

Mixture of Gaussians

- There are unknown clusters: $S_1, \ldots, S_k$.
- The points in S_i are distributed $MVN(\mu_i, \Sigma_i)$
- Each point x_i originates from cluster S_j with probability p_j.


25
20
15
10
5
0
−5
−10
−15
−15
−10
−5
0
5
10
15
20
25
30



$$S_1 \sim MVN\left((10,10), \begin{pmatrix} 29.25 & 13.5 \\ 13.5 & 20.25 \end{pmatrix}\right)$$

$$S_2 \sim MVN\left((0,0), \begin{pmatrix} 9 & -3.3 \\ -3.3 & 18 \end{pmatrix}\right)$$

$$p_1 = 0.25, p_2 = 0.75$$

In one dimension

- There are unknown clusters: $S_1, \ldots, S_k$.
- The points in S_i are distributed $N(\mu_i, \sigma_i^2)$
- Each point x_i originates from cluster S_j with probability p_j.

$$f_j(x) = \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{(x-\mu_j)^2}{2\sigma_j^2}}$$

$$L((\vec{p}, \vec{\mu}, \vec{\sigma}); \vec{x}) = \prod_{i=1}^{n} \sum_{j=1}^{k} p_j f_j(x_i)$$

For every i, we choose: $a_{ij} \geq 0, \sum_j a_{ij} = 1$

$$\begin{aligned}
\log L((\vec{p}, \vec{\mu}, \vec{\sigma}); \vec{x}) &= \sum_{i=1}^{n} \log \left(\sum_{j=1}^{k} p_j f_j(x_i) \right) \\
&= \sum_{i=1}^{n} \log \left(\sum_{j=1}^{k} a_{ij} \frac{p_j f_j(x_i)}{a_{ij}} \right) \\
&\geq \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \log(p_j f_j(x_i)) - a_{ij} \log(a_{ij})
\end{aligned}$$

The Expectation-Maximization Algorithm

- Start with a guess: (μ_i^0, σ_i^0)
- In each iteration t+1 set:

$$a_{ij} = Pr(x_i \in S_j \mid \vec{p}^t, \vec{\mu}^t, \vec{\sigma}^t) = \frac{p_j^t f_j^t(x_i)}{\sum_{m=1}^k p_m^t f_m^t(x_i)}$$

$$(\vec{p}^{t+1}, \vec{\mu}^{t+1}, \vec{\sigma}^{t+1}) = \arg\max_{\vec{\mu},\vec{\sigma},\vec{p}} \sum_{i=1}^n \sum_{j=1}^n a_{ij} \log(p_j f_j(x_i))$$

The Expectation-Maximization Algorithm

- Start with a guess: (μ_i^0, σ_i^0)
- In each iteration t+1 set:

$$a_{ij} = Pr(x_i \in S_j \mid \vec{p}^t, \vec{\mu}^t, \vec{\sigma}^t) = \frac{p_j^t f_j^t(x_i)}{\sum_{m=1}^k p_m^t f_m^t(x_i)}$$

$$p_j^{t+1} = \frac{\sum_{i=1}^n a_{ij}}{n}$$

$$\mu_j^{t+1} = \frac{\sum_{i=1}^n a_{ij} x_i}{\sum_{i=1}^n a_{ij}}$$

$$\sigma_j^{t+1} = \frac{\sum_{i=1}^n a_{ij} (x_i - \mu_j^{t+1})^2}{\sum_{i=1}^n a_{ij}}$$

The Expectation-Maximization Algorithm

$$g(\vec{p}, \vec{\mu}, \vec{\sigma}) := \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \log(p_j f_j(x_i)) - a_{ij} \log(a_{ij})$$

By construction:

$$\log L((\vec{p}, \vec{\mu}, \vec{\sigma}); \vec{x}) \geq g(\vec{p}, \vec{\mu}, \vec{\sigma})$$

$$\log L((\vec{p}^t, \vec{\mu}^t, \vec{\sigma}^t); \vec{x}) = g(\vec{p}^t, \vec{\mu}^t, \vec{\sigma}^t)$$

$$\begin{aligned}
\log L((\vec{p}^{t+1}, \vec{\mu}^{t+1}, \vec{\sigma}^{t+1}); \vec{x}) &\geq g(\vec{p}^{t+1}, \vec{\mu}^{t+1}, \vec{\sigma}^{t+1}) \\
&\geq g(\vec{p}^t, \vec{\mu}^t, \vec{\sigma}^t) \\
&= \log L((\vec{p}^t, \vec{\mu}^t, \vec{\sigma}^t); \vec{x})
\end{aligned}$$

The Expectation-Maximization Algorithm

Conclusion: The likelihood is non-decreasing in each iteration.

Stopping rule: When the likelihood flattens.

$$\begin{aligned}
\log L((\vec{p}^{t+1}, \vec{\mu}^{t+1}, \vec{\sigma}^{t+1}); \vec{x}) &\geq g(\vec{p}^{t+1}, \vec{\mu}^{t+1}, \vec{\sigma}^{t+1}) \\
&\geq g(\vec{p}^{t}, \vec{\mu}^{t}, \vec{\sigma}^{t}) \\
&= \log L((\vec{p}^{t}, \vec{\mu}^{t}, \vec{\sigma}^{t}); \vec{x})
\end{aligned}$$

Expectation Maximization (EM)

- D – given data
- θ – parameters that need to be estimated
- Z – missing (latent) variables

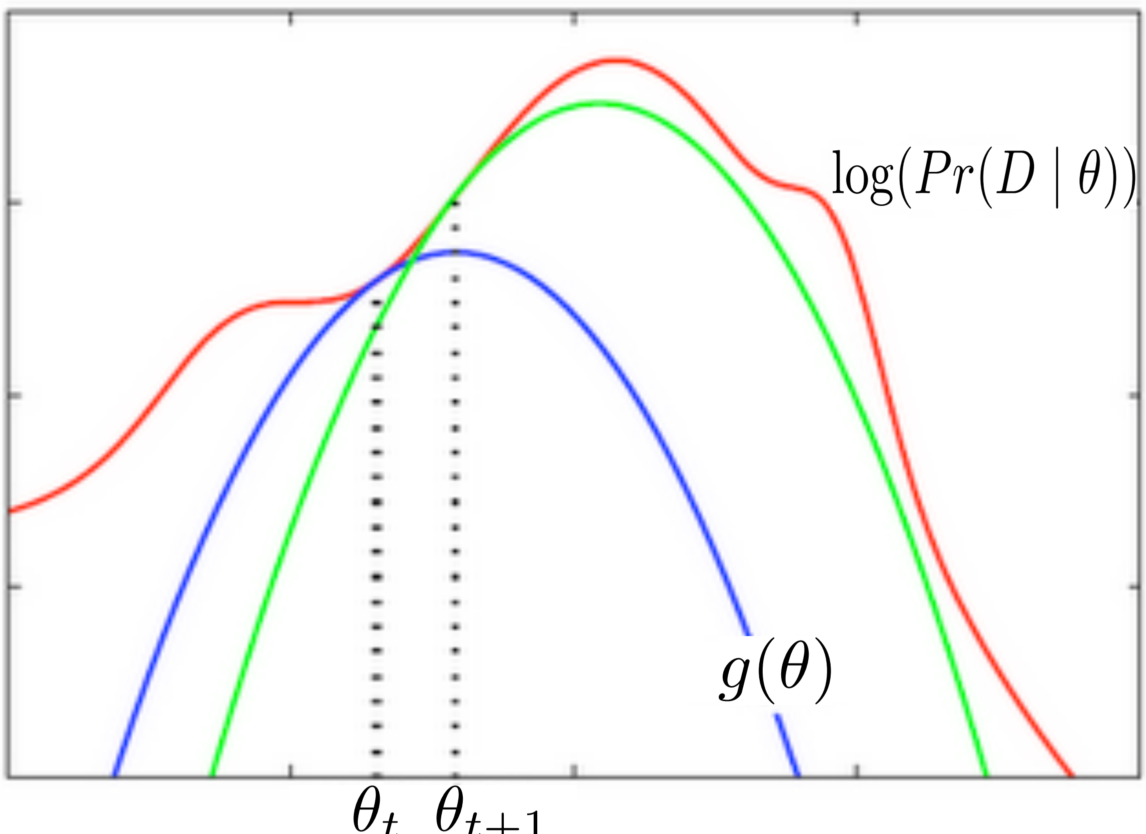
1. **E-step:** $Q(\theta \mid \theta_t) = E_{Z|D,\theta_t}[\log(Pr(D, Z \mid \theta)]$
2. **M-step:** $\theta_{t+1} := \arg\max_{\theta} Q(\theta \mid \theta_t)$

$$
\begin{aligned}
\log Pr(D \mid \theta) &= \log\left(\sum_z Pr(D, z \mid \theta)\right) \\
&= \log\left(\sum_z a_z \frac{Pr(D, z \mid \theta)}{a_z}\right) \\
&\geq \sum_z a_z \log\left(Pr(D, z \mid \theta)\right) - \sum_z a_z \log(a_z) \\
&= Q(\theta \mid \theta_t) - constant
\end{aligned}
$$

$$
\begin{aligned}
\log(Pr(D \mid \theta_{t+1})) &\geq Q(\theta_{t+1} \mid \theta_t) - constant \\
&\geq Q(\theta_t \mid \theta_t) - constant = \log(Pr(D \mid \theta_t))
\end{aligned}
$$

$$g(\theta) = Q(\theta \mid \theta_t) - \sum_z a_z \log(a_z)$$

$$\log Pr(D \mid \theta_{t+1}) \geq g(\theta_{t+1}) \geq g(\theta_t) = \log Pr(D \mid \theta_t)$$



EM - Comments

- No guarantee of optimization to local maximum.
- No guarantee of running times. Often it takes many iterations to converge.
- Efficiency: no matrix inversion is needed (e.g., in Newton). Generalized EM – no need to find the max in the M-step.
- Easy to implement.
- Numerical stability.
- Monotone – it is easy to ensure correctness in EM.
- Interpretation – provides interpretation for the latent variables.