Introduction to Machine Learning

Expectation Maximization

Andres Mendez-Vazquez

February 15, 2023

Outline

Cinvestav

Outline

Cinvestav

Maximum-Likelihood

We have a density function $p(\boldsymbol{x}|\Theta)$

Assume that we have a data set of size N, $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\}$

- This data is known as **evidence**.

We assume in addition that

The vectors are independent and identically distributed (i.i.d.) with distribution p under parameter θ.

Maximum-Likelihood

We have a density function $p(\boldsymbol{x}|\Theta)$

Assume that we have a data set of size N, $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\}$

- This data is known as **evidence**.

We assume in addition that

The vectors are independent and identically distributed (i.i.d.) with distribution p under parameter θ.

What Can We Do With The Evidence?

We may use the Bayes' Rule to estimate the parameters θ

$$p(\Theta|\mathcal{X}) = \frac{P(\mathcal{X}|\Theta)\,P(\Theta)}{P(\mathcal{X})} \quad (1)$$

Or, given a new observation $\tilde{x}$

$$p(\tilde{x}|\mathcal{X}) \quad (2)$$

I.e. to compute the probability of the new observation being supported by the evidence $\mathcal{X}$.

Thus

The former represents parameter estimation and the latter data prediction.

What Can We Do With The Evidence?

We may use the Bayes' Rule to estimate the parameters θ

$$p\left(\Theta|\mathcal{X}\right)=\frac{P\left(\mathcal{X}|\Theta\right)P\left(\Theta\right)}{P\left(\mathcal{X}\right)} \tag{1}$$

Or, given a new observation $\tilde{\boldsymbol{x}}$

$$p\left(\tilde{\boldsymbol{x}}|\mathcal{X}\right) \tag{2}$$

I.e. to compute the probability of the new observation being supported by the evidence $\mathcal{X}$.

Thus

The former represents parameter estimation and the latter data prediction.

What Can We Do With The Evidence?

We may use the Bayes' Rule to estimate the parameters θ

$$p(\Theta|\mathcal{X}) = \frac{P(\mathcal{X}|\Theta)\,P(\Theta)}{P(\mathcal{X})} \tag{1}$$

Or, given a new observation $\tilde{\boldsymbol{x}}$

$$p(\tilde{\boldsymbol{x}}|\mathcal{X}) \tag{2}$$

I.e. to compute the probability of the new observation being supported by the evidence $\mathcal{X}$.

Thus

The former represents parameter estimation and the latter data prediction.

Cinvestav

Focusing First on the Estimation of the Parameters θ

We can interpret the Bayes' Rule

$$p\left(\Theta|\mathcal{X}\right) = \frac{P\left(\mathcal{X}|\Theta\right)P\left(\Theta\right)}{P\left(\mathcal{X}\right)} \tag{3}$$

Interpreted as

$$posterior = \frac{likelihood \times prior}{evidence} \tag{4}$$

Thus, we want

$$likelihood = P\left(\mathcal{X}|\Theta\right)$$

Cinvestav

Focusing First on the Estimation of the Parameters θ

We can interpret the Bayes' Rule

$$p\left(\Theta|\mathcal{X}\right)=\frac{P\left(\mathcal{X}|\Theta\right)P\left(\Theta\right)}{P\left(\mathcal{X}\right)} \tag{3}$$

Interpreted as

$$posterior=\frac{likelihood\times prior}{evidence} \tag{4}$$

Thus, we want

$$likelihood=P\left(\mathcal{X}|\Theta\right)$$

Focusing First on the Estimation of the Parameters θ

We can interpret the Bayes' Rule

$$p(\Theta|\mathcal{X}) = \frac{P(\mathcal{X}|\Theta)\,P(\Theta)}{P(\mathcal{X})} \tag{3}$$

Interpreted as

$$posterior = \frac{likelihood \times prior}{evidence} \tag{4}$$

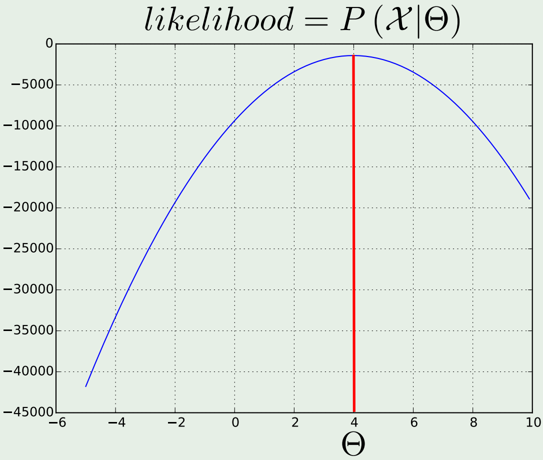
Thus, we want

$$likelihood = P(\mathcal{X}|\Theta)$$

Cinvestav

What we want...

We want to maximize the likelihood as a function of θ



Maximum-Likelihood

We have

$$p\left(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N | \Theta\right) = \prod_{i=1}^{N} p\left(\boldsymbol{x}_i | \Theta\right) \tag{5}$$

Also known as the likelihood function.

Because multiplication of quantities $p\left(\boldsymbol{x}_i | \Theta\right) \leq 1$ can be problematic

$$\mathcal{L}\left(\Theta | \mathcal{X}\right) = \log \prod_{i=1}^{N} p\left(\boldsymbol{x}_i | \Theta\right) = \sum_{i=1}^{N} \log p\left(\boldsymbol{x}_i | \Theta\right) \tag{6}$$

Maximum-Likelihood

We have

$$p\left(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N|\Theta\right) = \prod_{i=1}^{N} p\left(\boldsymbol{x}_i|\Theta\right) \quad (5)$$

Also known as the likelihood function.

Because multiplication of quantities $p\left(\boldsymbol{x}_i|\Theta\right) \leq 1$ can be problematic

$$\mathcal{L}\left(\Theta|\mathcal{X}\right) = \log \prod_{i=1}^{N} p\left(\boldsymbol{x}_i|\Theta\right) = \sum_{i=1}^{N} \log p\left(\boldsymbol{x}_i|\Theta\right) \quad (6)$$

Maximum-Likelihood

We want to find a Θ^*

$$\Theta^* = \text{argmax}_{\Theta} \mathcal{L}\left(\Theta|\mathcal{X}\right) \tag{7}$$

The classic method

$$\frac{\partial \mathcal{L}\left(\Theta|\mathcal{X}\right)}{\partial \theta_i} = 0 \ \forall \theta_i \in \Theta \tag{8}$$

Maximum-Likelihood

We want to find a Θ^*

$$\Theta^* = \text{argmax}_{\Theta} \mathcal{L}\left(\Theta|\mathcal{X}\right) \tag{7}$$

The classic method

$$\frac{\partial \mathcal{L}\left(\Theta|\mathcal{X}\right)}{\partial \theta_i} = 0 \; \forall \theta_i \in \Theta \tag{8}$$

What happened if we have incomplete data

Data could have been split

1. $\mathcal{X}$ = observed data or **incomplete** data
2. $\mathcal{Y}$ = unobserved data

For this type of problems

We have the famous Expectation Maximization (EM)

What happened if we have incomplete data

Data could have been split

1. $\mathcal{X}$ = observed data or **incomplete** data
2. $\mathcal{Y}$ = unobserved data

For this type of problems

We have the famous Expectation Maximization (EM)

Cinvestav

What happened if we have incomplete data

Data could have been split

1. $\mathcal{X}$ = observed data or **incomplete** data
2. $\mathcal{Y}$ = unobserved data

For this type of problems

We have the famous Expectation Maximization (EM)

Outline

Cinvestav

The Expectation Maximization

The EM algorithm

It was first developed by Dempster et al. (1977).

Its popularity comes from the fact

It can estimate an underlying distribution when data is incomplete or has missing values.

Two main applications

1. When missing values exists.
2. When a likelihood function can be simplified by assuming extra parameters that are **missing** or **hidden.**

The Expectation Maximization

The EM algorithm

It was first developed by Dempster et al. (1977).

Its popularity comes from the fact

It can estimate an underlying distribution when data is incomplete or has missing values.

Two main applications

1. When missing values exists.
2. When a likelihood function can be simplified by assuming extra parameters that are **missing** or **hidden.**

Cinvestav

The Expectation Maximization

The EM algorithm

It was first developed by Dempster et al. (1977).

Its popularity comes from the fact

It can estimate an underlying distribution when data is incomplete or has missing values.

Two main applications

1. When missing values exists.
2. When a likelihood function can be simplified by assuming extra parameters that are **missing** or **hidden**.

Cinvestav

The Expectation Maximization

The EM algorithm

It was first developed by Dempster et al. (1977).

Its popularity comes from the fact

It can estimate an underlying distribution when data is incomplete or has missing values.

Two main applications

1. When missing values exists.
2. When a likelihood function can be simplified by assuming extra parameters that are **missing** or **hidden.**

Cinvestav

Outline

Clustering

Given a series of data sets

Given the fact that Radial Gaussian Functions are Universal Approximators

.

- Samples $\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\}$ are the visible parameters
- The Gaussian distributions generating each of the samples are the hidden parameters

Then, we model the cluster as a mixture of Gaussian's

Cinvestav

Clustering

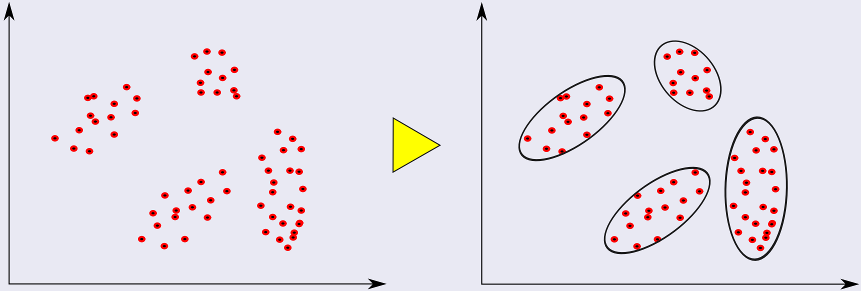
Given a series of data sets

Given the fact that Radial Gaussian Functions are Universal Approximators

.

- Samples $\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\}$ are the visible parameters
- The Gaussian distributions generating each of the samples are the hidden parameters

Then, we model the cluster as a mixture of Gaussian's



Natural Language Processing

Unsupervised induction of probabilistic context-free grammars

Here given a series of words $o_1, o_2, o_3, \ldots$ and normalized Context-Free Grammar

- We want to know the probabilities of each rule $P(i \rightarrow jk)$

Thus

- Here the you have two variables:
  - The Visible Ones: The sequence of words
  - The Hidden Ones: The rule that produces the possible sequence $o_i \rightarrow o_j$

Natural Language Processing

Unsupervised induction of probabilistic context-free grammars

Here given a series of words $o_1, o_2, o_3, ...$ and normalized Context-Free Grammar

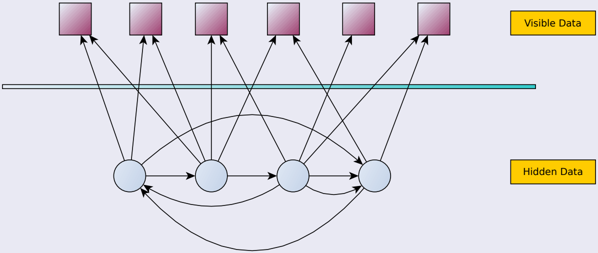
- We want to know the probabilities of each rule $P(i \to jk)$

Thus

- Here the you have two variables:
  - The Visible Ones: The sequence of words
  - The Hidden Ones: The rule that produces the possible sequence $o_i \to o_j$

Natural Language Processing

Baum-Welch Algorithm for Hidden Markov Models

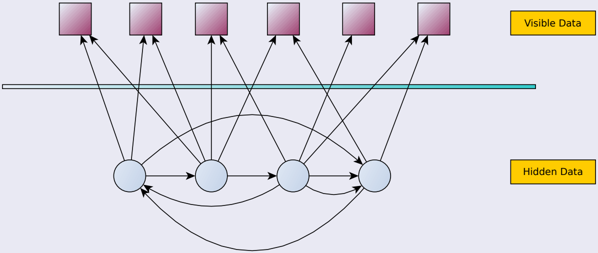


Here

- Hidden Variables: The circular nodes producing the data
- Visible Variables: The square nodes representing the samples.

Natural Language Processing

Baum-Welch Algorithm for Hidden Markov Models



Here

- Hidden Variables: The circular nodes producing the data
- Visible Variables: The square nodes representing the samples.

Outline

Cinvestav

Incomplete Data

We assume the following

Two parts of data

1. $\mathcal{X}$ = observed data or **incomplete** data
2. $\mathcal{Y}$ = unobserved data

Thus

$$\mathcal{Z} = (\mathcal{X}, \mathcal{Y}) = \text{Complete Data} \tag{9}$$

Thus, we have the following probability

$$p(z|\Theta) = p(x, y|\Theta) = p(y|x, \Theta)\, p(x|\Theta) \tag{10}$$

Cinvestav

Incomplete Data

We assume the following

Two parts of data

1. $\mathcal{X}$ = observed data or **incomplete** data
2. $\mathcal{Y}$ = unobserved data

Thus

$$\mathcal{Z} = (\mathcal{X}, \mathcal{Y}) = \text{Complete Data} \tag{9}$$

Thus, we have the following probability

$$p(z|\Theta) = p(x, y|\Theta) = p(y|x, \Theta)\, p(x|\Theta) \tag{10}$$

Incomplete Data

We assume the following

Two parts of data

1. $\mathcal{X}$ = observed data or **incomplete** data
2. $\mathcal{Y}$ = unobserved data

Thus

$$\mathcal{Z} = (\mathcal{X},\mathcal{Y}) = \text{Complete Data} \tag{9}$$

Thus, we have the following probability

$$p\left(\boldsymbol{z}|\Theta\right) = p\left(\boldsymbol{x},\boldsymbol{y}|\Theta\right) = p\left(\boldsymbol{y}|\boldsymbol{x},\Theta\right)p\left(\boldsymbol{x}|\Theta\right) \tag{10}$$

Incomplete Data

We assume the following

Two parts of data

1. $\mathcal{X}$ = observed data or **incomplete** data
2. $\mathcal{Y}$ = unobserved data

Thus

$$\mathcal{Z} = (\mathcal{X},\mathcal{Y}) = \text{Complete Data} \tag{9}$$

Thus, we have the following probability

$$p\left(\boldsymbol{z}|\Theta\right) = p\left(\boldsymbol{x},\boldsymbol{y}|\Theta\right) = p\left(\boldsymbol{y}|\boldsymbol{x},\Theta\right) p\left(\boldsymbol{x}|\Theta\right) \tag{10}$$

Cinvestav

Incomplete Data

We assume the following

Two parts of data

1. $\mathcal{X}$ = observed data or **incomplete** data
2. $\mathcal{Y}$ = unobserved data

Thus

$$\mathcal{Z} = (\mathcal{X},\mathcal{Y}) = \text{Complete Data} \tag{9}$$

Thus, we have the following probability

$$p(\boldsymbol{z}|\Theta) = p(\boldsymbol{x},\boldsymbol{y}|\Theta) = p(\boldsymbol{y}|\boldsymbol{x},\Theta)\,p(\boldsymbol{x}|\Theta) \tag{10}$$

New Likelihood Function

The New Likelihood Function

$$\mathcal{L}(\Theta|\mathcal{Z}) = \mathcal{L}(\Theta|\mathcal{X},\mathcal{Y}) = p(\mathcal{X},\mathcal{Y}|\Theta) \tag{11}$$

Note: The complete data likelihood.

New Likelihood Function

The New Likelihood Function

$$\mathcal{L}\left(\Theta|\mathcal{Z}\right)=\mathcal{L}\left(\Theta|\mathcal{X},\mathcal{Y}\right)=p\left(\mathcal{X},\mathcal{Y}|\Theta\right) \tag{11}$$

Note: The complete data likelihood.

Thus, we have

$$\mathcal{L}\left(\Theta|\mathcal{X},\mathcal{Y}\right)=p\left(\mathcal{X},\mathcal{Y}|\Theta\right)=p\left(\mathcal{Y}|\mathcal{X},\Theta\right)p\left(\mathcal{X}|\Theta\right) \tag{12}$$

Did you notice?

- $p\left(\mathcal{X}|\Theta\right)$ is the likelihood of the observed data.
- $p\left(\mathcal{Y}|\mathcal{X},\Theta\right)$ is the likelihood of the no-observed data under the observed data!!!

New Likelihood Function

The New Likelihood Function

$$\mathcal{L}(\Theta|\mathcal{Z}) = \mathcal{L}(\Theta|\mathcal{X},\mathcal{Y}) = p(\mathcal{X},\mathcal{Y}|\Theta) \quad (11)$$

Note: The complete data likelihood.

Thus, we have

$$\mathcal{L}(\Theta|\mathcal{X},\mathcal{Y}) = p(\mathcal{X},\mathcal{Y}|\Theta) = p(\mathcal{Y}|\mathcal{X},\Theta)\, p(\mathcal{X}|\Theta) \quad (12)$$

Did you notice?

- $p(\mathcal{X}|\Theta)$ is the likelihood of the observed data.
- $p(\mathcal{Y}|\mathcal{X},\Theta)$ is the likelihood of the no-observed data under the observed data!!!

Rewriting

This can be rewritten as

$$\mathcal{L}(\Theta|\mathcal{X},\mathcal{Y}) = h_{\mathcal{X},\Theta}(\mathcal{Y}) \quad (13)$$

This basically signify that $\mathcal{X}, \Theta$ are constant and the only random part is $\mathcal{Y}$.

In addition

$$\mathcal{L}(\Theta|\mathcal{X}) \quad (14)$$

It is known as the incomplete-data likelihood function.

Rewriting

This can be rewritten as

$$\mathcal{L}(\Theta|\mathcal{X},\mathcal{Y}) = h_{\mathcal{X},\Theta}(\mathcal{Y}) \tag{13}$$

This basically signify that $\mathcal{X},\Theta$ are constant and the only random part is $\mathcal{Y}$.

In addition

$$\mathcal{L}(\Theta|\mathcal{X}) \tag{14}$$

It is known as the incomplete-data likelihood function.

Cinvestav

Thus

We can connect both incomplete-complete data equations by doing the following

$$
\begin{aligned}
\mathcal{L}(\Theta|\mathcal{X}) =& p(\mathcal{X}|\Theta) \\
=& \sum_{\mathcal{Y}} p(\mathcal{X},\mathcal{Y}|\Theta) \\
=& \sum_{\mathcal{Y}} p(\mathcal{Y}|\mathcal{X},\Theta)\, p(\mathcal{X}|\Theta) \\
=& \sum \left(\prod_{i=1}^{N} p(x_i|\Theta)\right)_{\mathcal{Y}} p(\mathcal{Y}|\mathcal{X},\Theta)
\end{aligned}
$$

Cinvestav

Thus

We can connect both incomplete-complete data equations by doing the following

$$
\begin{aligned}
\mathcal{L}(\Theta|\mathcal{X}) =& p(\mathcal{X}|\Theta) \\
=& \sum_{\mathcal{Y}} p(\mathcal{X},\mathcal{Y}|\Theta) \\
=& \sum_{\mathcal{Y}} p(\mathcal{Y}|\mathcal{X},\Theta)\, p(\mathcal{X}|\Theta) \\
=& \sum \left(\prod_{i=1}^{N} p(x_i|\Theta) \right)_{\mathcal{Y}} p(\mathcal{Y}|\mathcal{X},\Theta)
\end{aligned}
$$

Cinvestav

Thus

We can connect both incomplete-complete data equations by doing the following

$$
\begin{aligned}
\mathcal{L}(\Theta|\mathcal{X}) =& p(\mathcal{X}|\Theta) \\
=& \sum_{\mathcal{Y}} p(\mathcal{X},\mathcal{Y}|\Theta) \\
=& \sum_{\mathcal{Y}} p(\mathcal{Y}|\mathcal{X},\Theta)\, p(\mathcal{X}|\Theta) \\
=& \sum \left(\prod_{i=1}^{N} p(x_i|\Theta) \right)_{\mathcal{Y}} p(\mathcal{Y}|\mathcal{X},\Theta)
\end{aligned}
$$

Cinvestav

Thus

We can connect both incomplete-complete data equations by doing the following

$$
\begin{aligned}
\mathcal{L}(\Theta|\mathcal{X}) =& p(\mathcal{X}|\Theta)\\
=& \sum_{\mathcal{Y}} p(\mathcal{X},\mathcal{Y}|\Theta)\\
=& \sum_{\mathcal{Y}} p(\mathcal{Y}|\mathcal{X},\Theta)\, p(\mathcal{X}|\Theta)\\
=& \sum \left(\prod_{i=1}^{N} p(x_i|\Theta)\right)_{\mathcal{Y}} p(\mathcal{Y}|\mathcal{X},\Theta)
\end{aligned}
$$

Cinvestav

Remarks

Problems

Normally, it is almost impossible to obtain a closed analytical solution for the previous equation.

However

We can use the expected value of $\log p(\mathcal{X}, \mathcal{Y}|\Theta)$, which allows us to find an iterative procedure to approximate the solution.

Remarks

Problems

Normally, it is almost impossible to obtain a closed analytical solution for the previous equation.

However

We can use the expected value of $\log p(\mathcal{X}, \mathcal{Y}|\Theta)$, which allows us to find an iterative procedure to approximate the solution.

The function we would like to have

The Q function

We want an estimation of the complete-data log-likelihood

$$\log p\left(\mathcal{X}, \mathcal{Y}|\Theta\right) \tag{15}$$

Based in the info provided by $\mathcal{X}, \Theta_{n-1}$ where Θ_{n-1} is a previously estimated set of parameters at step n.

Think about the following, if we want to remove $\mathcal{Y}$

$$\int \left[\log p\left(\mathcal{X}, \mathcal{Y}|\Theta\right)\right] p\left(\mathcal{Y}|\mathcal{X}, \Theta_{n-1}\right) d\mathcal{Y} \tag{16}$$

Remark: We integrate out $\mathcal{Y}$ - Actually, this is the expected value of $\log p\left(\mathcal{X}, \mathcal{Y}|\Theta\right)$.

The function we would like to have

The Q function

We want an estimation of the complete-data log-likelihood

$$\log p(\mathcal{X}, \mathcal{Y}|\Theta) \tag{15}$$

Based in the info provided by $\mathcal{X}, \Theta_{n-1}$ where Θ_{n-1} is a previously estimated set of parameters at step n.

Think about the following, if we want to remove $\mathcal{Y}$

$$\int [\log p(\mathcal{X}, \mathcal{Y}|\Theta)]\, p(\mathcal{Y}|\mathcal{X}, \Theta_{n-1})\, d\mathcal{Y} \tag{16}$$

Remark: We integrate out $\mathcal{Y}$ - Actually, this is the expected value of $\log p(\mathcal{X}, \mathcal{Y}|\Theta)$.

Outline

Cinvestav

Use the Expected Value

Then, we want an iterative method to guess Θ from Θ_{n-1}

$$Q\left(\Theta,\Theta_{n-1}\right)=E\left[\log p\left(\mathcal{X},\mathcal{Y}|\Theta\right)|\mathcal{X},\Theta_{n-1}\right] \tag{17}$$

Take in account that

1. $\mathcal{X},\Theta_{n-1}$ are taken as constants.
2. Θ is a normal variable that we wish to adjust.
3. $\mathcal{Y}$ is a random variable governed by distribution $p\left(\mathcal{Y}|\mathcal{X},\Theta_{n-1}\right)$=marginal distribution of missing data.

Use the Expected Value

Then, we want an iterative method to guess Θ from Θ_{n-1}

$$Q\left(\Theta, \Theta_{n-1}\right) = E\left[\log p\left(\mathcal{X}, \mathcal{Y} | \Theta\right) | \mathcal{X}, \Theta_{n-1}\right] \tag{17}$$

Take in account that

1. $\mathcal{X}, \Theta_{n-1}$ are taken as constants.
2. Θ is a normal variable that we wish to adjust.
3. $\mathcal{Y}$ is a random variable governed by distribution $p\left(\mathcal{Y}|\mathcal{X}, \Theta_{n-1}\right)$=marginal distribution of missing data.

Cinvestav

Use the Expected Value

Then, we want an iterative method to guess Θ from Θ_{n-1}

$$Q\left(\Theta, \Theta_{n-1}\right) = E\left[\log p\left(\mathcal{X}, \mathcal{Y} | \Theta\right) | \mathcal{X}, \Theta_{n-1}\right] \tag{17}$$

Take in account that

1. $\mathcal{X}, \Theta_{n-1}$ are taken as constants.
2. Θ is a normal variable that we wish to adjust.
3. $\mathcal{Y}$ is a random variable governed by distribution $p\left(\mathcal{Y}|\mathcal{X}, \Theta_{n-1}\right)$=marginal distribution of missing data.

Use the Expected Value

Then, we want an iterative method to guess Θ from Θ_{n-1}

$$Q\left(\Theta, \Theta_{n-1}\right) = E\left[\log p\left(\mathcal{X}, \mathcal{Y}|\Theta\right)|\mathcal{X}, \Theta_{n-1}\right] \tag{17}$$

Take in account that

1. $\mathcal{X}, \Theta_{n-1}$ are taken as constants.
2. Θ is a normal variable that we wish to adjust.
3. $\mathcal{Y}$ is a random variable governed by distribution $p\left(\mathcal{Y}|\mathcal{X}, \Theta_{n-1}\right)$=marginal distribution of missing data.

Another Interpretation

Given the previous information

$$E\left[\log p\left(\mathcal{X},\mathcal{Y}|\Theta\right)|\mathcal{X},\Theta_{n-1}\right]=\int_{\mathcal{Y}\in\mathbb{Y}}\log p\left(\mathcal{X},\mathcal{Y}|\Theta\right)p\left(\mathcal{Y}|\mathcal{X},\Theta_{n-1}\right)d\mathcal{Y}$$

Something Notable

1. In the best of cases, this marginal distribution is a simple analytical expression of the assumed parameter Θ_{n-1}.
2. In the worst of cases, this density might be very hard to obtain.

Actually, we use

$$p\left(\mathcal{Y},\mathcal{X}|\Theta_{n-1}\right)=p\left(\mathcal{Y}|\mathcal{X},\Theta_{n-1}\right)p\left(\mathcal{X}|\Theta_{n-1}\right)\qquad(18)$$

which is not dependent on Θ.

Another Interpretation

Given the previous information

$$E\left[\log p\left(\mathcal{X},\mathcal{Y}|\Theta\right)|\mathcal{X},\Theta_{n-1}\right]=\int_{\mathcal{Y}\in\mathbb{Y}}\log p\left(\mathcal{X},\mathcal{Y}|\Theta\right)p\left(\mathcal{Y}|\mathcal{X},\Theta_{n-1}\right)d\mathcal{Y}$$

Something Notable

1. In the best of cases, this marginal distribution is a simple analytical expression of the assumed parameter Θ_{n-1}.
2. In the worst of cases, this density might be very hard to obtain.

Actually, we use

$$p\left(\mathcal{Y},\mathcal{X}|\Theta_{n-1}\right)=p\left(\mathcal{Y}|\mathcal{X},\Theta_{n-1}\right)p\left(\mathcal{X}|\Theta_{n-1}\right) \quad (18)$$

which is not dependent on Θ.

Another Interpretation

Given the previous information

$$E\left[\log p\left(\mathcal{X},\mathcal{Y}|\Theta\right)|\mathcal{X},\Theta_{n-1}\right]=\int_{\boldsymbol{\mathcal{Y}}\in\mathbb{Y}}\log p\left(\mathcal{X},\mathcal{Y}|\Theta\right)p\left(\mathcal{Y}|\mathcal{X},\Theta_{n-1}\right)d\mathcal{Y}$$

Something Notable

1. In the best of cases, this marginal distribution is a simple analytical expression of the assumed parameter Θ_{n-1}.
2. In the worst of cases, this density might be very hard to obtain.

Actually, we use

$$p\left(\mathcal{Y},\mathcal{X}|\Theta_{n-1}\right)=p\left(\mathcal{Y}|\mathcal{X},\Theta_{n-1}\right)p\left(\mathcal{X}|\Theta_{n-1}\right) \tag{18}$$

which is not dependent on Θ.

Another Interpretation

Given the previous information

$$E\left[\log p\left(\mathcal{X},\mathcal{Y}|\Theta\right)|\mathcal{X},\Theta_{n-1}\right]=\int_{\mathcal{Y}\in\mathbb{Y}}\log p\left(\mathcal{X},\mathcal{Y}|\Theta\right)p\left(\mathcal{Y}|\mathcal{X},\Theta_{n-1}\right)d\mathcal{Y}$$

Something Notable

1. In the best of cases, this marginal distribution is a simple analytical expression of the assumed parameter Θ_{n-1}.
2. In the worst of cases, this density might be very hard to obtain.

Actually, we use

$$p\left(\mathcal{Y},\mathcal{X}|\Theta_{n-1}\right)=p\left(\mathcal{Y}|\mathcal{X},\Theta_{n-1}\right)p\left(\mathcal{X}|\Theta_{n-1}\right) \tag{18}$$

which is not dependent on Θ.

Cinvestav

Outline

Cinvestav

Back to the Q function

The intuition

We have the following analogy:

- Consider $h(\theta, Y)$ a function
  - θ a constant
  - $Y \sim p_Y(y)$, a random variable with distribution $p_Y(y)$.

Thus, if Y is a discrete random variable

$$q(\theta) = E_Y[h(\theta, Y)] = \sum_y h(\theta, y)\, p_Y(y) \tag{19}$$

Back to the Q function

The intuition

We have the following analogy:

- Consider $h(\theta, \boldsymbol{Y})$ a function
  - θ a constant
  - $\boldsymbol{Y} \sim p_{\boldsymbol{Y}}(y)$, a random variable with distribution $p_{\boldsymbol{Y}}(y)$.

Thus, if $\boldsymbol{Y}$ is a discrete random variable

$$q(\theta) = E_{\boldsymbol{Y}}\left[h(\theta, \boldsymbol{Y})\right] = \sum_{y} h(\theta, y)\, p_{\boldsymbol{Y}}(y) \quad (19)$$

Cinvestav

Back to the Q function

The intuition

We have the following analogy:

- Consider $h(\theta, \boldsymbol{Y})$ a function
  - θ a constant
  - $\boldsymbol{Y} \sim p_{\boldsymbol{Y}}(y)$, a random variable with distribution $p_{\boldsymbol{Y}}(y)$.

Thus, if $\boldsymbol{Y}$ is a discrete random variable

$$q(\theta) = E_{\boldsymbol{Y}}\left[h(\theta, \boldsymbol{Y})\right] = \sum_{y} h(\theta, y)\, p_{\boldsymbol{Y}}(y) \quad (19)$$

Cinvestav

Back to the Q function

The intuition

We have the following analogy:

- Consider $h(\theta, \boldsymbol{Y})$ a function
  - θ a constant
  - $\boldsymbol{Y} \sim p_{\mathbf{Y}}(y)$, a random variable with distribution $p_{\mathbf{Y}}(y)$.

Thus, if $\boldsymbol{Y}$ is a discrete random variable

$$q(\theta) = E_{\mathbf{Y}}[h(\theta, \boldsymbol{Y})] = \sum_{y} h(\theta, y) p_{\mathbf{Y}}(y) \tag{19}$$

Back to the Q function

The intuition

We have the following analogy:

- Consider $h(\theta, \boldsymbol{Y})$ a function
  - θ a constant
  - $\boldsymbol{Y} \sim p_{\boldsymbol{Y}}(y)$, a random variable with distribution $p_{\boldsymbol{Y}}(y)$.

Thus, if $\boldsymbol{Y}$ is a discrete random variable

$$q(\theta) = E_{\boldsymbol{Y}}\left[h(\theta, \boldsymbol{Y})\right] = \sum_{y} h(\theta, y)\, p_{\boldsymbol{Y}}(y) \tag{19}$$

Cinvestav

Why E-step!!!

From here the name

This is basically the E-step

The second step

It tries to maximize the Q function

$$\Theta_n = \text{argmax}_{\Theta} Q\left(\Theta, \Theta_{n-1}\right) \quad (20)$$

Cinvestav

Why E-step!!!

From here the name

This is basically the E-step

The second step

It tries to maximize the Q function

$$\Theta_n = \text{argmax}_{\Theta} Q\left(\Theta, \Theta_{n-1}\right) \quad (20)$$

Cinvestav

Why E-step!!!

From here the name

This is basically the E-step

The second step

It tries to maximize the Q function

$$\Theta_n = \text{argmax}_{\Theta} Q\left(\Theta, \Theta_{n-1}\right) \tag{20}$$

Derivation of the EM-Algorithm

The likelihood function we are going to use

Let $\mathcal{X}$ be a random vector which results from a parametrized family:

$$\mathcal{L}(\Theta) = \ln \mathcal{P}(\mathcal{X}|\Theta) \tag{21}$$

Note: $\ln(x)$ is a strictly increasing function.

We wish to compute Θ

Based on an estimate Θ_n (After the n^{th}) such that $\mathcal{L}(\Theta) > \mathcal{L}(\Theta_n)$

Or the maximization of the difference

$$\mathcal{L}(\Theta) - \mathcal{L}(\Theta_n) = \ln \mathcal{P}(\mathcal{X}|\Theta) - \ln \mathcal{P}(\mathcal{X}|\Theta_n) \tag{22}$$

Derivation of the EM-Algorithm

The likelihood function we are going to use

Let $\mathcal{X}$ be a random vector which results from a parametrized family:

$$\mathcal{L}(\Theta) = \ln \mathcal{P}(\mathcal{X}|\Theta) \tag{21}$$

Note: $\ln(x)$ is a strictly increasing function.

We wish to compute Θ

Based on an estimate Θ_n (After the n^{th}) such that $\mathcal{L}(\Theta) > \mathcal{L}(\Theta_n)$

Or the maximization of the difference

$$\mathcal{L}(\Theta) - \mathcal{L}(\Theta_n) = \ln \mathcal{P}(\mathcal{X}|\Theta) - \ln \mathcal{P}(\mathcal{X}|\Theta_n) \tag{22}$$

Derivation of the EM-Algorithm

The likelihood function we are going to use

Let $\mathcal{X}$ be a random vector which results from a parametrized family:

$$\mathcal{L}(\Theta) = \ln \mathcal{P}(\mathcal{X}|\Theta) \tag{21}$$

Note: $\ln(x)$ is a strictly increasing function.

We wish to compute Θ

Based on an estimate Θ_n (After the n^{th}) such that $\mathcal{L}(\Theta) > \mathcal{L}(\Theta_n)$

Or the maximization of the difference

$$\mathcal{L}(\Theta) - \mathcal{L}(\Theta_n) = \ln \mathcal{P}(\mathcal{X}|\Theta) - \ln \mathcal{P}(\mathcal{X}|\Theta_n) \tag{22}$$

Cinvestav

Outline

Cinvestav

Introducing the Hidden Features

Given that the hidden random vector $\mathcal{Y}$ exits with y values

$$\mathcal{P}\left(\mathcal{X}|\Theta\right) = \sum_{y} \mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right) \tag{23}$$

Thus, using our first constraint $\mathcal{L}\left(\Theta\right) - \mathcal{L}\left(\Theta_n\right)$

$$\mathcal{L}\left(\Theta\right) - \mathcal{L}\left(\Theta_n\right) = \ln\left(\sum_{y} \mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right)\right) - \ln \mathcal{P}\left(\mathcal{X}|\Theta_n\right) \tag{24}$$

Cinvestav

Introducing the Hidden Features

Given that the hidden random vector $\mathcal{Y}$ exits with y values

$$\mathcal{P}\left(\mathcal{X}|\Theta\right) = \sum_{y} \mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right) \quad (23)$$

Thus, using our first constraint $\mathcal{L}\left(\Theta\right) - \mathcal{L}\left(\Theta_n\right)$

$$\mathcal{L}\left(\Theta\right) - \mathcal{L}\left(\Theta_n\right) = \ln\left(\sum_{y} \mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right)\right) - \ln \mathcal{P}\left(\mathcal{X}|\Theta_n\right) \quad (24)$$

Cinvestav

Here, we introduce some concepts of convexity

For Convexity

Theorem (Jensen's inequality)

Let f be a convex function defined on an interval I. If $x_1, x_2, ..., x_n \in I$ and $\lambda_1, \lambda_2, ..., \lambda_n \geq 0$ with $\sum_{i=1}^{n} \lambda_i = 1$, then

$$f\left(\sum_{i=1}^{n} \lambda_i x_i\right) \leq \sum_{i=1}^{n} \lambda_i f(x_i) \tag{25}$$

Proof:

For $n = 1$

We have the trivial case

For $n = 2$

The convexity definition.

Now the inductive hypothesis

We assume that the theorem is true for some n.

Proof:

For $n = 1$

We have the trivial case

For $n = 2$

The convexity definition.

Now the inductive hypothesis

We assume that the theorem is true for some n.

Cinvestav

Proof:

For $n = 1$

We have the trivial case

For $n = 2$

The convexity definition.

Now the inductive hypothesis

We assume that the theorem is true for some n.

Now, we have

The following linear combination for λ_i

$$f\left(\sum_{i=1}^{n+1}\lambda_i x_i\right) = f\left(\lambda_{n+1}x_{n+1} + \sum_{i=1}^{n}\lambda_i x_i\right)$$

$$= f\left(\lambda_{n+1}x_{n+1} + \frac{(1-\lambda_{n+1})}{(1-\lambda_{n+1})}\sum_{i=1}^{n}\lambda_i x_i\right)$$

$$\leq \lambda_{n+1}f\left(x_{n+1}\right) + \left(1-\lambda_{n+1}\right)f\left(\frac{1}{(1-\lambda_{n+1})}\sum_{i=1}^{n}\lambda_i x_i\right)$$

Now, we have

The following linear combination for λ_i

$$
\begin{aligned}
f\left(\sum_{i=1}^{n+1} \lambda_i x_i\right) &= f\left(\lambda_{n+1} x_{n+1} + \sum_{i=1}^{n} \lambda_i x_i\right) \\
&= f\left(\lambda_{n+1} x_{n+1} + \frac{(1-\lambda_{n+1})}{(1-\lambda_{n+1})} \sum_{i=1}^{n} \lambda_i x_i\right) \\
&\leq \lambda_{n+1} f\left(x_{n+1}\right) + \left(1-\lambda_{n+1}\right) f\left(\frac{1}{(1-\lambda_{n+1})} \sum_{i=1}^{n} \lambda_i x_i\right)
\end{aligned}
$$

Now, we have

The following linear combination for λ_i

$$
\begin{aligned}
f\left(\sum_{i=1}^{n+1}\lambda_i x_i\right) &= f\left(\lambda_{n+1}x_{n+1}+\sum_{i=1}^{n}\lambda_i x_i\right)\\
&= f\left(\lambda_{n+1}x_{n+1}+\frac{(1-\lambda_{n+1})}{(1-\lambda_{n+1})}\sum_{i=1}^{n}\lambda_i x_i\right)\\
&\leq \lambda_{n+1}f\left(x_{n+1}\right)+\left(1-\lambda_{n+1}\right)f\left(\frac{1}{(1-\lambda_{n+1})}\sum_{i=1}^{n}\lambda_i x_i\right)
\end{aligned}
$$

Cinvestav

Did you notice?

Something Notable

$$\sum_{i=1}^{n+1} \lambda_i = 1$$

Thus

$$\sum_{i=1}^{n} \lambda_i = 1 - \lambda_{n+1}$$

Finally

$$\frac{1}{(1 - \lambda_{n+1})} \sum_{i=1}^{n} \lambda_i = 1$$

Cinvestav

Did you notice?

Something Notable

$$\sum_{i=1}^{n+1} \lambda_i = 1$$

Thus

$$\sum_{i=1}^{n} \lambda_i = 1 - \lambda_{n+1}$$

Finally

$$\frac{1}{(1-\lambda_{n+1})} \sum_{i=1}^{n} \lambda_i = 1$$

Cinvestav

Did you notice?

Something Notable

$$\sum_{i=1}^{n+1} \lambda_i = 1$$

Thus

$$\sum_{i=1}^{n} \lambda_i = 1 - \lambda_{n+1}$$

Finally

$$\frac{1}{(1-\lambda_{n+1})} \sum_{i=1}^{n} \lambda_i = 1$$

Now

We have that

$$f\left(\sum_{i=1}^{n+1}\lambda_i x_i\right) \leq \lambda_{n+1} f\left(x_{n+1}\right) + \left(1-\lambda_{n+1}\right) f\left(\frac{1}{\left(1-\lambda_{n+1}\right)}\sum_{i=1}^{n}\lambda_i x_i\right)$$

$$\leq \lambda_{n+1} f\left(x_{n+1}\right) + \left(1-\lambda_{n+1}\right) \frac{1}{\left(1-\lambda_{n+1}\right)}\sum_{i=1}^{n}\lambda_i f\left(x_i\right)$$

$$\leq \lambda_{n+1} f\left(x_{n+1}\right) + \sum_{i=1}^{n}\lambda_i f\left(x_i\right) \text{ Q.E.D.}$$

Now

We have that

$$
\begin{aligned}
f\left(\sum_{i=1}^{n+1} \lambda_i x_i\right) &\leq \lambda_{n+1} f\left(x_{n+1}\right) + \left(1-\lambda_{n+1}\right) f\left(\frac{1}{\left(1-\lambda_{n+1}\right)} \sum_{i=1}^{n} \lambda_i x_i\right) \\
&\leq \lambda_{n+1} f\left(x_{n+1}\right) + \left(1-\lambda_{n+1}\right) \frac{1}{\left(1-\lambda_{n+1}\right)} \sum_{i=1}^{n} \lambda_i f\left(x_i\right) \\
&\leq \lambda_{n+1} f\left(x_{n+1}\right) + \sum_{i=1}^{n} \lambda_i f\left(x_i\right) \text{ Q.E.D.}
\end{aligned}
$$

Now

We have that

$$
\begin{aligned}
f\left(\sum_{i=1}^{n+1} \lambda_i x_i\right) &\leq \lambda_{n+1} f\left(x_{n+1}\right) + \left(1-\lambda_{n+1}\right) f\left(\frac{1}{\left(1-\lambda_{n+1}\right)} \sum_{i=1}^{n} \lambda_i x_i\right) \\
&\leq \lambda_{n+1} f\left(x_{n+1}\right) + \left(1-\lambda_{n+1}\right) \frac{1}{\left(1-\lambda_{n+1}\right)} \sum_{i=1}^{n} \lambda_i f\left(x_i\right) \\
&\leq \lambda_{n+1} f\left(x_{n+1}\right) + \sum_{i=1}^{n} \lambda_i f\left(x_i\right) \text{ Q.E.D.}
\end{aligned}
$$

Cinvestav

Thus, for concave functions

It is possible to shown that

Given $\ln(x)$ a concave function:

$$\ln\left[\sum_{i=1}^{n}\lambda_i x_i\right] \geq \sum_{i=1}^{n}\lambda_i \ln(x_i)$$

If we take in consideration

Assume that the $\lambda_i = \mathcal{P}(y|\mathcal{X},\Theta_n)$. We know that

1. $\mathcal{P}(y|\mathcal{X},\Theta_n) \geq 0$
2. $\sum_y \mathcal{P}(y|\mathcal{X},\Theta_n) = 1$

Thus, for concave functions

It is possible to shown that

Given $\ln(x)$ a concave function:

$$\ln\left[\sum_{i=1}^{n}\lambda_i x_i\right] \geq \sum_{i=1}^{n}\lambda_i \ln(x_i)$$

If we take in consideration

Assume that the $\lambda_i = \mathcal{P}(y|\mathcal{X},\Theta_n)$. We know that

1. $\mathcal{P}(y|\mathcal{X},\Theta_n) \geq 0$
2. $\sum_y \mathcal{P}(y|\mathcal{X},\Theta_n) = 1$

Cinvestav

We have

First

$$\mathcal{L}(\Theta) - \mathcal{L}(\Theta_n) = \ln\left(\sum_{y} \mathcal{P}(\mathcal{X}|y,\Theta)\,\mathcal{P}(y|\Theta)\right) - \ln \mathcal{P}(\mathcal{X}|\Theta_n)$$

$$= \ln\left(\sum_{y} \mathcal{P}(\mathcal{X}|y,\Theta)\,\mathcal{P}(y|\Theta)\,\frac{\mathcal{P}(y|\mathcal{X},\Theta_n)}{\mathcal{P}(y|\mathcal{X},\Theta_n)}\right) - \ln \mathcal{P}(\mathcal{X}|\Theta_n)$$

$$= \ln\left(\sum_{y} \mathcal{P}(y|\mathcal{X},\Theta_n)\,\frac{\mathcal{P}(\mathcal{X}|y,\Theta)\,\mathcal{P}(y|\Theta)}{\mathcal{P}(y|\mathcal{X},\Theta_n)}\right) - \ln \mathcal{P}(\mathcal{X}|\Theta_n)$$

$$\geq \sum_{y} \mathcal{P}(y|\mathcal{X},\Theta_n) \ln\left(\frac{\mathcal{P}(\mathcal{X}|y,\Theta)\,\mathcal{P}(y|\Theta)}{\mathcal{P}(y|\mathcal{X},\Theta_n)}\right) - \cdots$$

$\sum_{y} \mathcal{P}(y|\mathcal{X},\Theta_n) \ln \mathcal{P}(\mathcal{X}|\Theta_n)$ Why this?

We have

First

$$
\begin{aligned}
\mathcal{L}(\Theta) - \mathcal{L}(\Theta_n) &= \ln\left(\sum_y \mathcal{P}(\mathcal{X}|y,\Theta)\,\mathcal{P}(y|\Theta)\right) - \ln \mathcal{P}(\mathcal{X}|\Theta_n) \\
&= \ln\left(\sum_y \mathcal{P}(\mathcal{X}|y,\Theta)\,\mathcal{P}(y|\Theta)\,\frac{\mathcal{P}(y|\mathcal{X},\Theta_n)}{\mathcal{P}(y|\mathcal{X},\Theta_n)}\right) - \ln \mathcal{P}(\mathcal{X}|\Theta_n) \\
&= \ln\left(\sum_y \mathcal{P}(y|\mathcal{X},\Theta_n)\,\frac{\mathcal{P}(\mathcal{X}|y,\Theta)\,\mathcal{P}(y|\Theta)}{\mathcal{P}(y|\mathcal{X},\Theta_n)}\right) - \ln \mathcal{P}(\mathcal{X}|\Theta_n) \\
&\geq \sum_y \mathcal{P}(y|\mathcal{X},\Theta_n)\ln\left(\frac{\mathcal{P}(\mathcal{X}|y,\Theta)\,\mathcal{P}(y|\Theta)}{\mathcal{P}(y|\mathcal{X},\Theta_n)}\right) - \cdots
\end{aligned}
$$

$\sum_y \mathcal{P}(y|\mathcal{X},\Theta_n) \ln \mathcal{P}(\mathcal{X}|\Theta_n)$ Why this?

We have

First

$$
\begin{aligned}
\mathcal{L}(\Theta) - \mathcal{L}(\Theta_n) &= \ln\left(\sum_y \mathcal{P}(\mathcal{X}|y,\Theta)\,\mathcal{P}(y|\Theta)\right) - \ln \mathcal{P}(\mathcal{X}|\Theta_n) \\
&= \ln\left(\sum_y \mathcal{P}(\mathcal{X}|y,\Theta)\,\mathcal{P}(y|\Theta)\,\frac{\mathcal{P}(y|\mathcal{X},\Theta_n)}{\mathcal{P}(y|\mathcal{X},\Theta_n)}\right) - \ln \mathcal{P}(\mathcal{X}|\Theta_n) \\
&= \ln\left(\sum_y \mathcal{P}(y|\mathcal{X},\Theta_n)\,\frac{\mathcal{P}(\mathcal{X}|y,\Theta)\,\mathcal{P}(y|\Theta)}{\mathcal{P}(y|\mathcal{X},\Theta_n)}\right) - \ln \mathcal{P}(\mathcal{X}|\Theta_n) \\
&\geq \sum_y \mathcal{P}(y|\mathcal{X},\Theta_n) \ln\left(\frac{\mathcal{P}(\mathcal{X}|y,\Theta)\,\mathcal{P}(y|\Theta)}{\mathcal{P}(y|\mathcal{X},\Theta_n)}\right) - \cdots
\end{aligned}
$$

$\sum_y \mathcal{P}(y|\mathcal{X},\Theta_n) \ln \mathcal{P}(\mathcal{X}|\Theta_n)$ Why this?

We have

First

$$
\begin{aligned}
\mathcal{L}(\Theta) - \mathcal{L}(\Theta_n) &= \ln \left(\sum_y \mathcal{P}(\mathcal{X}|y,\Theta)\, \mathcal{P}(y|\Theta) \right) - \ln \mathcal{P}(\mathcal{X}|\Theta_n) \\
&= \ln \left(\sum_y \mathcal{P}(\mathcal{X}|y,\Theta)\, \mathcal{P}(y|\Theta) \frac{\mathcal{P}(y|\mathcal{X},\Theta_n)}{\mathcal{P}(y|\mathcal{X},\Theta_n)} \right) - \ln \mathcal{P}(\mathcal{X}|\Theta_n) \\
&= \ln \left(\sum_y \mathcal{P}(y|\mathcal{X},\Theta_n) \frac{\mathcal{P}(\mathcal{X}|y,\Theta)\, \mathcal{P}(y|\Theta)}{\mathcal{P}(y|\mathcal{X},\Theta_n)} \right) - \ln \mathcal{P}(\mathcal{X}|\Theta_n) \\
&\geq \sum_y \mathcal{P}(y|\mathcal{X},\Theta_n) \ln \left(\frac{\mathcal{P}(\mathcal{X}|y,\Theta)\, \mathcal{P}(y|\Theta)}{\mathcal{P}(y|\mathcal{X},\Theta_n)} \right) - \dots \\
& \sum_y \mathcal{P}(y|\mathcal{X},\Theta_n) \ln \mathcal{P}(\mathcal{X}|\Theta_n) \text{ Why this?}
\end{aligned}
$$

Cinvestav

Next

Because

$$\sum_{y} \mathcal{P}\left(y|\mathcal{X}, \Theta_n\right) = 1$$

Then

$$\mathcal{L}\left(\Theta\right) - \mathcal{L}\left(\Theta_n\right) \geq \sum_{y} \mathcal{P}\left(y|\mathcal{X}, \Theta_n\right) \ln\left(\frac{\mathcal{P}\left(\mathcal{X}|y, \Theta\right)\mathcal{P}\left(y|\Theta\right)}{\mathcal{P}\left(y|\mathcal{X}, \Theta_n\right)\mathcal{P}\left(\mathcal{X}|\Theta_n\right)}\right)$$

Next

Because

$$\sum_{y} \mathcal{P}\left(y|\mathcal{X}, \Theta_n\right) = 1$$

Then

$$\begin{aligned}\mathcal{L}\left(\Theta\right) - \mathcal{L}\left(\Theta_n\right) \geq & \sum_{y} \mathcal{P}\left(y|\mathcal{X}, \Theta_n\right) \ln\left(\frac{\mathcal{P}\left(\mathcal{X}|y, \Theta\right)\mathcal{P}\left(y|\Theta\right)}{\mathcal{P}\left(y|\mathcal{X}, \Theta_n\right)\mathcal{P}\left(\mathcal{X}|\Theta_n\right)}\right)\\ =&\Delta\left(\Theta|\Theta_n\right)\end{aligned}$$

Then, we have

Then, we have proved that

$$\mathcal{L}(\Theta) \geq \mathcal{L}(\Theta_n) + \Delta(\Theta|\Theta_n) \tag{26}$$

Then, we define a new function

$$l(\Theta|\Theta_n) = \mathcal{L}(\Theta_n) + \Delta(\Theta|\Theta_n) \tag{27}$$

Thus $l(\Theta|\Theta_n)$

It is bounded from above by $\mathcal{L}(\Theta)$ i.e $l(\Theta|\Theta_n) \leq \mathcal{L}(\Theta)$

Then, we have

Then, we have proved that

$$\mathcal{L}(\Theta) \geq \mathcal{L}(\Theta_n) + \Delta(\Theta|\Theta_n) \tag{26}$$

Then, we define a new function

$$l(\Theta|\Theta_n) = \mathcal{L}(\Theta_n) + \Delta(\Theta|\Theta_n) \tag{27}$$

Thus $l(\Theta|\Theta_n)$

It is bounded from above by $\mathcal{L}(\Theta)$ i.e $l(\Theta|\Theta_n) \leq \mathcal{L}(\Theta)$

Then, we have

Then, we have proved that

$$\mathcal{L}(\Theta) \geq \mathcal{L}(\Theta_n) + \Delta(\Theta|\Theta_n) \tag{26}$$

Then, we define a new function

$$l(\Theta|\Theta_n) = \mathcal{L}(\Theta_n) + \Delta(\Theta|\Theta_n) \tag{27}$$

Thus $l(\Theta|\Theta_n)$

It is bounded from above by $\mathcal{L}(\Theta)$ i.e $l(\Theta|\Theta_n) \leq \mathcal{L}(\Theta)$

Now, we can do the following

We evaluate in Θ_n

$$
\begin{aligned}
l\left(\Theta_n|\Theta_n\right) =& \mathcal{L}\left(\Theta_n\right) + \Delta\left(\Theta_n|\Theta_n\right) \\
=& \mathcal{L}\left(\Theta_n\right) + \sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\frac{\mathcal{P}\left(\mathcal{X}|y,\Theta_n\right)\mathcal{P}\left(y|\Theta_n\right)}{\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\mathcal{P}\left(\mathcal{X}|\Theta_n\right)}\right) \\
=& \mathcal{L}\left(\Theta_n\right) + \sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\frac{\mathcal{P}\left(\mathcal{X},y|\Theta_n\right)}{\mathcal{P}\left(\mathcal{X},y|\Theta_n\right)}\right) \\
=& \mathcal{L}\left(\Theta_n\right)
\end{aligned}
$$

This means that

For $\Theta = \Theta_n$, functions $\mathcal{L}\left(\Theta\right)$ and $l\left(\Theta|\Theta_n\right)$ are equal

Now, we can do the following

We evaluate in Θ_n

$$
\begin{aligned}
l\left(\Theta_n|\Theta_n\right) =& \mathcal{L}\left(\Theta_n\right) + \Delta\left(\Theta_n|\Theta_n\right)\\
=& \mathcal{L}\left(\Theta_n\right) + \sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\ln\left(\frac{\mathcal{P}\left(\mathcal{X}|y,\Theta_n\right)\mathcal{P}\left(y|\Theta_n\right)}{\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\mathcal{P}\left(\mathcal{X}|\Theta_n\right)}\right)\\
=& \mathcal{L}\left(\Theta_n\right) + \sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\ln\left(\frac{\mathcal{P}\left(\mathcal{X},y|\Theta_n\right)}{\mathcal{P}\left(\mathcal{X},y|\Theta_n\right)}\right)\\
=& \mathcal{L}\left(\Theta_n\right)
\end{aligned}
$$

This means that

For $\Theta = \Theta_n$, functions $\mathcal{L}\left(\Theta\right)$ and $l\left(\Theta|\Theta_n\right)$ are equal

Cinvestav

Now, we can do the following

We evaluate in Θ_n

$$
\begin{aligned}
l(\Theta_n|\Theta_n) =& \mathcal{L}(\Theta_n) + \Delta(\Theta_n|\Theta_n)\\
=& \mathcal{L}(\Theta_n) + \sum_y \mathcal{P}(y|\mathcal{X},\Theta_n) \ln\left(\frac{\mathcal{P}(\mathcal{X}|y,\Theta_n)\,\mathcal{P}(y|\Theta_n)}{\mathcal{P}(y|\mathcal{X},\Theta_n)\,\mathcal{P}(\mathcal{X}|\Theta_n)}\right)\\
=& \mathcal{L}(\Theta_n) + \sum_y \mathcal{P}(y|\mathcal{X},\Theta_n) \ln\left(\frac{\mathcal{P}(\mathcal{X},y|\Theta_n)}{\mathcal{P}(\mathcal{X},y|\Theta_n)}\right)\\
=& \mathcal{L}(\Theta_n)
\end{aligned}
$$

This means that

For $\Theta = \Theta_n$, functions $\mathcal{L}(\Theta)$ and $l(\Theta|\Theta_n)$ are equal

Cinvestav

Now, we can do the following

We evaluate in Θ_n

$$
\begin{aligned}
l\left(\Theta_n|\Theta_n\right) =& \mathcal{L}\left(\Theta_n\right)+\Delta\left(\Theta_n|\Theta_n\right)\\
=& \mathcal{L}\left(\Theta_n\right)+\sum_{y}\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\ln\left(\frac{\mathcal{P}\left(\mathcal{X}|y,\Theta_n\right)\mathcal{P}\left(y|\Theta_n\right)}{\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\mathcal{P}\left(\mathcal{X}|\Theta_n\right)}\right)\\
=& \mathcal{L}\left(\Theta_n\right)+\sum_{y}\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\ln\left(\frac{\mathcal{P}\left(\mathcal{X},y|\Theta_n\right)}{\mathcal{P}\left(\mathcal{X},y|\Theta_n\right)}\right)\\
=& \mathcal{L}\left(\Theta_n\right)
\end{aligned}
$$

This means that

For $\Theta=\Theta_n$, functions $\mathcal{L}\left(\Theta\right)$ and $l\left(\Theta|\Theta_n\right)$ are equal

Now, we can do the following

We evaluate in Θ_n

$$
\begin{aligned}
l\left(\Theta_n|\Theta_n\right) =& \mathcal{L}\left(\Theta_n\right) + \Delta\left(\Theta_n|\Theta_n\right)\\
=& \mathcal{L}\left(\Theta_n\right) + \sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\ln\left(\frac{\mathcal{P}\left(\mathcal{X}|y,\Theta_n\right)\mathcal{P}\left(y|\Theta_n\right)}{\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\mathcal{P}\left(\mathcal{X}|\Theta_n\right)}\right)\\
=& \mathcal{L}\left(\Theta_n\right) + \sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\ln\left(\frac{\mathcal{P}\left(\mathcal{X},y|\Theta_n\right)}{\mathcal{P}\left(\mathcal{X},y|\Theta_n\right)}\right)\\
=& \mathcal{L}\left(\Theta_n\right)
\end{aligned}
$$

This means that

For $\Theta = \Theta_n$, functions $\mathcal{L}\left(\Theta\right)$ and $l\left(\Theta|\Theta_n\right)$ are equal

Cinvestav

Therefore

The function $l\left(\Theta|\Theta_n\right)$ has the following properties

1. It is bounded from above by $\mathcal{L}\left(\Theta\right)$ i.e $l\left(\Theta|\Theta_n\right) \leq \mathcal{L}\left(\Theta\right)$.
2. For $\Theta = \Theta_n$, functions $\mathcal{L}\left(\Theta\right)$ and $l\left(\Theta|\Theta_n\right)$ are equal.
3. The function $l\left(\Theta|\Theta_n\right)$ is concave... How?

Therefore

The function $l\left(\Theta|\Theta_n\right)$ has the following properties

1. It is bounded from above by $\mathcal{L}\left(\Theta\right)$ i.e $l\left(\Theta|\Theta_n\right) \leq \mathcal{L}\left(\Theta\right)$.
2. For $\Theta = \Theta_n$, functions $\mathcal{L}\left(\Theta\right)$ and $l\left(\Theta|\Theta_n\right)$ are equal.
3. The function $l\left(\Theta|\Theta_n\right)$ is concave... How?

Therefore

The function $l\left(\Theta|\Theta_n\right)$ has the following properties

1. It is bounded from above by $\mathcal{L}\left(\Theta\right)$ i.e $l\left(\Theta|\Theta_n\right) \leq \mathcal{L}\left(\Theta\right)$.
2. For $\Theta = \Theta_n$, functions $\mathcal{L}\left(\Theta\right)$ and $l\left(\Theta|\Theta_n\right)$ are equal.
3. The function $l\left(\Theta|\Theta_n\right)$ is concave... How?

Outline

Cinvestav

First

We have the value $\mathcal{L}(\Theta_n)$

We know that $\mathcal{L}(\Theta_n)$ is constant i.e. an offset value

What about $\Delta(\Theta|\Theta_n)$

$$\sum_{y} \mathcal{P}(y|\mathcal{X},\Theta_n) \ln \left(\frac{\mathcal{P}(\mathcal{X}|y,\Theta)\,\mathcal{P}(y|\Theta)}{\mathcal{P}(y|\mathcal{X},\Theta_n)\,\mathcal{P}(\mathcal{X}|\Theta_n)} \right)$$

We have that the ln is a concave function

$$\ln \left(\frac{\mathcal{P}(\mathcal{X}|y,\Theta)\,\mathcal{P}(y|\Theta)}{\mathcal{P}(y|\mathcal{X},\Theta_n)\,\mathcal{P}(\mathcal{X}|\Theta_n)} \right)$$

First

We have the value $\mathcal{L}(\Theta_n)$

We know that $\mathcal{L}(\Theta_n)$ is constant i.e. an offset value

What about $\Delta(\Theta|\Theta_n)$

$$\sum_y \mathcal{P}(y|\mathcal{X},\Theta_n)\ln\left(\frac{\mathcal{P}(\mathcal{X}|y,\Theta)\,\mathcal{P}(y|\Theta)}{\mathcal{P}(y|\mathcal{X},\Theta_n)\,\mathcal{P}(\mathcal{X}|\Theta_n)}\right)$$

We have that the ln is a concave function

$$\ln\left(\frac{\mathcal{P}(\mathcal{X}|y,\Theta)\,\mathcal{P}(y|\Theta)}{\mathcal{P}(y|\mathcal{X},\Theta_n)\,\mathcal{P}(\mathcal{X}|\Theta_n)}\right)$$

First

We have the value $\mathcal{L}(\Theta_n)$

We know that $\mathcal{L}(\Theta_n)$ is constant i.e. an offset value

What about $\Delta(\Theta|\Theta_n)$

$$\sum_{y} \mathcal{P}(y|\mathcal{X},\Theta_n) \ln\left(\frac{\mathcal{P}(\mathcal{X}|y,\Theta)\,\mathcal{P}(y|\Theta)}{\mathcal{P}(y|\mathcal{X},\Theta_n)\,\mathcal{P}(\mathcal{X}|\Theta_n)}\right)$$

We have that the $\ln$ is a concave function

$$\ln\left(\frac{\mathcal{P}(\mathcal{X}|y,\Theta)\,\mathcal{P}(y|\Theta)}{\mathcal{P}(y|\mathcal{X},\Theta_n)\,\mathcal{P}(\mathcal{X}|\Theta_n)}\right)$$

Cinvestav

Therefore

Each element is concave

$$\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\ln\left(\frac{\mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right)}{\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\mathcal{P}\left(\mathcal{X}|\Theta_n\right)}\right)$$

Therefore, the sum of concave functions is a concave function

$$\sum_{y}\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\ln\left(\frac{\mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right)}{\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\mathcal{P}\left(\mathcal{X}|\Theta_n\right)}\right)$$

Therefore

Each element is concave

$$\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\ln\left(\frac{\mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right)}{\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\mathcal{P}\left(\mathcal{X}|\Theta_n\right)}\right)$$

Therefore, the sum of concave functions is a concave function

$$\sum_{y}\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\ln\left(\frac{\mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right)}{\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\mathcal{P}\left(\mathcal{X}|\Theta_n\right)}\right)$$

Cinvestav

Outline

Given the Concave Function

Thus, we have that

1. We can select Θ_n such that $l\left(\Theta|\Theta_n\right)$ is maximized.
2. Thus, given a Θ_n, we can generate Θ_{n+1}.

The process can be seen in the following graph

Given the Concave Function

Thus, we have that

1. We can select Θ_n such that $l(\Theta|\Theta_n)$ is maximized.
2. Thus, given a Θ_n, we can generate Θ_{n+1}.

The process can be seen in the following graph

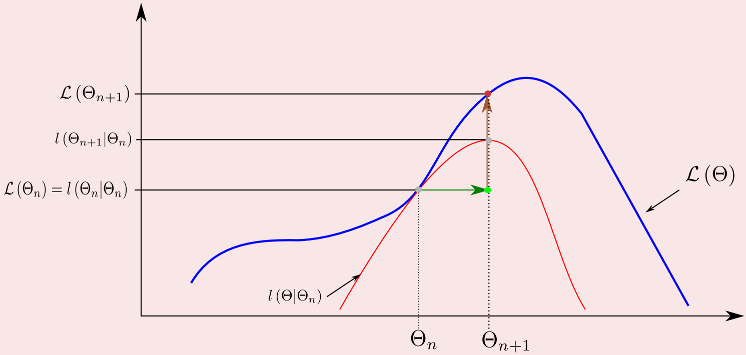
Cinvestav

Given the Concave Function

Thus, we have that

1. We can select Θ_n such that $l(\Theta|\Theta_n)$ is maximized.
2. Thus, given a Θ_n, we can generate Θ_{n+1}.

The process can be seen in the following graph



Given

The Previous Constraints

1. $l\left(\Theta|\Theta_n\right)$ is bounded from above by $\mathcal{L}\left(\Theta\right)$

$$l\left(\Theta|\Theta_n\right) \leq \mathcal{L}\left(\Theta\right)$$

2. For $\Theta = \Theta_n$, functions $\mathcal{L}\left(\Theta\right)$ and $l\left(\Theta|\Theta_n\right)$ are equal

$$\mathcal{L}\left(\Theta_n\right) = l\left(\Theta|\Theta_n\right)$$

3. The function $l\left(\Theta|\Theta_n\right)$ is concave

Given

The Previous Constraints

1. $l\left(\Theta|\Theta_n\right)$ is bounded from above by $\mathcal{L}\left(\Theta\right)$

$$l\left(\Theta|\Theta_n\right) \leq \mathcal{L}\left(\Theta\right)$$

2. For $\Theta = \Theta_n$, functions $\mathcal{L}\left(\Theta\right)$ and $l\left(\Theta|\Theta_n\right)$ are equal

$$\mathcal{L}\left(\Theta_n\right) = l\left(\Theta|\Theta_n\right)$$

3. The function $l\left(\Theta|\Theta_n\right)$ is concave

Cinvestav

Given

The Previous Constraints

1. $l\left(\Theta|\Theta_n\right)$ is bounded from above by $\mathcal{L}\left(\Theta\right)$

$$l\left(\Theta|\Theta_n\right) \leq \mathcal{L}\left(\Theta\right)$$

2. For $\Theta = \Theta_n$, functions $\mathcal{L}\left(\Theta\right)$ and $l\left(\Theta|\Theta_n\right)$ are equal

$$\mathcal{L}\left(\Theta_n\right) = l\left(\Theta|\Theta_n\right)$$

3. The function $l\left(\Theta|\Theta_n\right)$ is concave

Cinvestav

Outline

Cinvestav

From

The following

$$\begin{aligned}
\Theta_{n+1} =& \text{argmax}_{\Theta}\left\{l\left(\Theta|\Theta_n\right)\right\}\\
=& \text{argmax}_{\Theta}\left\{\mathcal{L}\left(\Theta_n\right)+\sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\ln\left(\frac{\mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right)}{\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\mathcal{P}\left(\mathcal{X}|\Theta_n\right)}\right)\right\}\\
& \text{The terms with } \Theta_n \text{ are constants.}\\
\approx& \text{argmax}_{\Theta}\left\{\sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\ln\left(\mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right)\right)\right\}\\
=& \text{argmax}_{\Theta}\left\{\sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\ln\left(\frac{\mathcal{P}\left(\mathcal{X},y|\Theta\right)}{\mathcal{P}\left(y|\Theta\right)}\frac{\mathcal{P}\left(y,\Theta\right)}{\mathcal{P}\left(\Theta\right)}\right)\right\}\\
=& \text{argmax}_{\Theta}\left\{\sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\ln\left(\frac{\frac{\mathcal{P}(\mathcal{X},y,\Theta)}{\mathcal{P}(\Theta)}}{\frac{\mathcal{P}(y,\Theta)}{\mathcal{P}(\Theta)}}\frac{\mathcal{P}\left(y,\Theta\right)}{\mathcal{P}\left(\Theta\right)}\right)\right\}
\end{aligned}$$

From

The following

$$
\begin{aligned}
\Theta_{n+1} = & \text{argmax}_{\Theta} \left\{ l\left(\Theta|\Theta_n\right)\right\} \\
= & \text{argmax}_{\Theta} \left\{ \mathcal{L}\left(\Theta_n\right) + \sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\frac{\mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right)}{\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\mathcal{P}\left(\mathcal{X}|\Theta_n\right)}\right)\right\}
\end{aligned}
$$

The terms with Θ_n are constants.

$$
\begin{aligned}
\approx & \text{argmax}_{\Theta} \left\{ \sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right)\right)\right\} \\
= & \text{argmax}_{\Theta} \left\{ \sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\frac{\mathcal{P}\left(\mathcal{X},y|\Theta\right)}{\mathcal{P}\left(y|\Theta\right)}\frac{\mathcal{P}\left(y,\Theta\right)}{\mathcal{P}\left(\Theta\right)}\right)\right\} \\
= & \text{argmax}_{\Theta} \left\{ \sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\frac{\frac{\mathcal{P}(\mathcal{X},y,\Theta)}{\mathcal{P}(\Theta)}}{\frac{\mathcal{P}(y,\Theta)}{\mathcal{P}(\Theta)}}\frac{\mathcal{P}\left(y,\Theta\right)}{\mathcal{P}\left(\Theta\right)}\right)\right\}
\end{aligned}
$$

From

The following

$$
\begin{aligned}
\Theta_{n+1} =& \text{argmax}_{\Theta} \left\{ l\left(\Theta | \Theta_n\right) \right\} \\
=& \text{argmax}_{\Theta} \left\{ \mathcal{L}\left(\Theta_n\right) + \sum_y \mathcal{P}\left(y | \mathcal{X}, \Theta_n\right) \ln \left(\frac{\mathcal{P}\left(\mathcal{X} | y, \Theta\right) \mathcal{P}\left(y | \Theta\right)}{\mathcal{P}\left(y | \mathcal{X}, \Theta_n\right) \mathcal{P}\left(\mathcal{X} | \Theta_n\right)} \right) \right\} \\
& \text{The terms with } \Theta_n \text{ are constants.} \\
\approx& \text{argmax}_{\Theta} \left\{ \sum_y \mathcal{P}\left(y | \mathcal{X}, \Theta_n\right) \ln \left(\mathcal{P}\left(\mathcal{X} | y, \Theta\right) \mathcal{P}\left(y | \Theta\right) \right) \right\} \\
=& \text{argmax}_{\Theta} \left\{ \sum_y \mathcal{P}\left(y | \mathcal{X}, \Theta_n\right) \ln \left(\frac{\mathcal{P}\left(\mathcal{X}, y | \Theta\right)}{\mathcal{P}\left(y | \Theta\right)} \frac{\mathcal{P}\left(y, \Theta\right)}{\mathcal{P}\left(\Theta\right)} \right) \right\} \\
=& \text{argmax}_{\Theta} \left\{ \sum_y \mathcal{P}\left(y | \mathcal{X}, \Theta_n\right) \ln \left(\frac{\frac{\mathcal{P}(\mathcal{X}, y, \Theta)}{\mathcal{P}(\Theta)}}{\frac{\mathcal{P}(y, \Theta)}{\mathcal{P}(\Theta)}} \frac{\mathcal{P}\left(y, \Theta\right)}{\mathcal{P}\left(\Theta\right)} \right) \right\}
\end{aligned}
$$

From

The following

$$
\begin{aligned}
\Theta_{n+1} &= \text{argmax}_{\Theta} \left\{ l\left(\Theta|\Theta_n\right) \right\} \\
&= \text{argmax}_{\Theta} \left\{ \mathcal{L}\left(\Theta_n\right) + \sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln \left(\frac{\mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right)}{\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\mathcal{P}\left(\mathcal{X}|\Theta_n\right)} \right) \right\}
\end{aligned}
$$

The terms with Θ_n are constants.

$$
\begin{aligned}
&\approx \text{argmax}_{\Theta} \left\{ \sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln \left(\mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right) \right) \right\} \\
&= \text{argmax}_{\Theta} \left\{ \sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln \left(\frac{\mathcal{P}\left(\mathcal{X},y|\Theta\right)}{\mathcal{P}\left(y|\Theta\right)} \frac{\mathcal{P}\left(y,\Theta\right)}{\mathcal{P}\left(\Theta\right)} \right) \right\} \\
&= \text{argmax}_{\Theta} \left\{ \sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln \left(\frac{\frac{\mathcal{P}(\mathcal{X},y,\Theta)}{\mathcal{P}(\Theta)}}{\frac{\mathcal{P}(y,\Theta)}{\mathcal{P}(\Theta)}} \frac{\mathcal{P}\left(y,\Theta\right)}{\mathcal{P}\left(\Theta\right)} \right) \right\}
\end{aligned}
$$

From

The following

$$
\begin{aligned}
\Theta_{n+1} = & \text{argmax}_{\Theta} \left\{ l\left(\Theta|\Theta_n\right) \right\} \\
= & \text{argmax}_{\Theta} \left\{ \mathcal{L}\left(\Theta_n\right) + \sum_y \mathcal{P}\left(y|\mathcal{X}, \Theta_n\right) \ln\left(\frac{\mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right)}{\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\mathcal{P}\left(\mathcal{X}|\Theta_n\right)} \right) \right\} \\
& \quad \text{The terms with } \Theta_n \text{ are constants.} \\
\approx & \text{argmax}_{\Theta} \left\{ \sum_y \mathcal{P}\left(y|\mathcal{X}, \Theta_n\right) \ln\left(\mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right) \right) \right\} \\
= & \text{argmax}_{\Theta} \left\{ \sum_y \mathcal{P}\left(y|\mathcal{X}, \Theta_n\right) \ln\left(\frac{\mathcal{P}\left(\mathcal{X},y|\Theta\right)}{\mathcal{P}\left(y|\Theta\right)} \frac{\mathcal{P}\left(y,\Theta\right)}{\mathcal{P}\left(\Theta\right)} \right) \right\} \\
= & \text{argmax}_{\Theta} \left\{ \sum_y \mathcal{P}\left(y|\mathcal{X}, \Theta_n\right) \ln\left(\frac{\frac{\mathcal{P}(\mathcal{X},y,\Theta)}{\mathcal{P}(\Theta)}}{\frac{\mathcal{P}(y,\Theta)}{\mathcal{P}(\Theta)}} \frac{\mathcal{P}\left(y,\Theta\right)}{\mathcal{P}\left(\Theta\right)} \right) \right\}
\end{aligned}
$$

Thus

Then

$$
\begin{aligned}
\theta_{n+1} &= \text{argmax}_{\Theta} \left\{ \sum_{y} \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\frac{\mathcal{P}\left(\mathcal{X},y,\Theta\right)}{\mathcal{P}\left(y,\Theta\right)} \frac{\mathcal{P}\left(y,\Theta\right)}{\mathcal{P}\left(\Theta\right)} \right) \right\} \\
&= \text{argmax}_{\Theta} \left\{ \sum_{y} \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\frac{\mathcal{P}\left(\mathcal{X},y,\Theta\right)}{\mathcal{P}\left(\Theta\right)} \right) \right\} \\
&= \text{argmax}_{\Theta} \left\{ \sum_{y} \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\mathcal{P}\left(\mathcal{X},y|\Theta\right) \right) \right\} \\
&= \text{argmax}_{\Theta} \left\{ E_{y|\mathcal{X},\Theta_n}\left[\ln\left(\mathcal{P}\left(\mathcal{X},y|\Theta\right) \right) \right] \right\}
\end{aligned}
$$

Then $\text{argmax}_{\Theta} \left\{ l\left(\Theta|\Theta_n\right) \right\} \approx \text{argmax}_{\Theta} \left\{ E_{y|\mathcal{X},\Theta_n}\left[\ln\left(\mathcal{P}\left(\mathcal{X},y|\Theta\right) \right) \right] \right\}$

Cinvestav

Thus

Then

$$
\begin{aligned}
\theta_{n+1} &= \text{argmax}_{\Theta} \left\{ \sum_{y} \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\frac{\mathcal{P}\left(\mathcal{X},y,\Theta\right)}{\mathcal{P}\left(y,\Theta\right)} \frac{\mathcal{P}\left(y,\Theta\right)}{\mathcal{P}\left(\Theta\right)} \right) \right\} \\
&= \text{argmax}_{\Theta} \left\{ \sum_{y} \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\frac{\mathcal{P}\left(\mathcal{X},y,\Theta\right)}{\mathcal{P}\left(\Theta\right)} \right) \right\} \\
&= \text{argmax}_{\Theta} \left\{ \sum_{y} \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\mathcal{P}\left(\mathcal{X},y|\Theta\right) \right) \right\} \\
&= \text{argmax}_{\Theta} \left\{ E_{y|\mathcal{X},\Theta_n}\left[\ln\left(\mathcal{P}\left(\mathcal{X},y|\Theta\right) \right) \right] \right\}
\end{aligned}
$$

Then $\text{argmax}_{\Theta} \left\{ l\left(\Theta|\Theta_n\right) \right\} \approx \text{argmax}_{\Theta} \left\{ E_{y|\mathcal{X},\Theta_n}\left[\ln\left(\mathcal{P}\left(\mathcal{X},y|\Theta\right) \right) \right] \right\}$

Thus

Then

$$
\begin{aligned}
\theta_{n+1} =& \text{argmax}_{\Theta} \left\{ \sum_{y} \mathcal{P}\left(y|\mathcal{X}, \Theta_n\right) \ln\left(\frac{\mathcal{P}\left(\mathcal{X}, y, \Theta\right)}{\mathcal{P}\left(y, \Theta\right)} \frac{\mathcal{P}\left(y, \Theta\right)}{\mathcal{P}\left(\Theta\right)} \right) \right\} \\
=& \text{argmax}_{\Theta} \left\{ \sum_{y} \mathcal{P}\left(y|\mathcal{X}, \Theta_n\right) \ln\left(\frac{\mathcal{P}\left(\mathcal{X}, y, \Theta\right)}{\mathcal{P}\left(\Theta\right)} \right) \right\} \\
=& \text{argmax}_{\Theta} \left\{ \sum_{y} \mathcal{P}\left(y|\mathcal{X}, \Theta_n\right) \ln\left(\mathcal{P}\left(\mathcal{X}, y|\Theta\right) \right) \right\} \\
=& \text{argmax}_{\Theta} \left\{ E_{y|\mathcal{X},\Theta_n}\left[\ln\left(\mathcal{P}\left(\mathcal{X}, y|\Theta\right) \right) \right] \right\}
\end{aligned}
$$

Then $\text{argmax}_{\Theta} \left\{ l\left(\Theta|\Theta_n\right) \right\} \approx \text{argmax}_{\Theta} \left\{ E_{y|\mathcal{X},\Theta_n}\left[\ln\left(\mathcal{P}\left(\mathcal{X}, y|\Theta\right) \right) \right] \right\}$

Thus

Then

$$
\begin{aligned}
\theta_{n+1} =& \text{argmax}_{\Theta} \left\{ \sum_{y} \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\frac{\mathcal{P}\left(\mathcal{X},y,\Theta\right)}{\mathcal{P}\left(y,\Theta\right)} \frac{\mathcal{P}\left(y,\Theta\right)}{\mathcal{P}\left(\Theta\right)} \right) \right\} \\
=& \text{argmax}_{\Theta} \left\{ \sum_{y} \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\frac{\mathcal{P}\left(\mathcal{X},y,\Theta\right)}{\mathcal{P}\left(\Theta\right)} \right) \right\} \\
=& \text{argmax}_{\Theta} \left\{ \sum_{y} \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\mathcal{P}\left(\mathcal{X},y|\Theta\right)\right) \right\} \\
=& \text{argmax}_{\Theta} \left\{ E_{y|\mathcal{X},\Theta_n}\left[\ln\left(\mathcal{P}\left(\mathcal{X},y|\Theta\right)\right)\right] \right\}
\end{aligned}
$$

Then $\text{argmax}_{\Theta}\left\{l\left(\Theta|\Theta_n\right)\right\} \approx \text{argmax}_{\Theta}\left\{E_{y|\mathcal{X},\Theta_n}\left[\ln\left(\mathcal{P}\left(\mathcal{X},y|\Theta\right)\right)\right]\right\}$

Cinvestav

Thus

Then

$$
\begin{aligned}
\theta_{n+1} =& \text{argmax}_{\Theta}\left\{\sum_{y}\mathcal{P}\left(y|\mathcal{X},\Theta_{n}\right)\ln\left(\frac{\mathcal{P}\left(\mathcal{X},y,\Theta\right)}{\mathcal{P}\left(y,\Theta\right)}\frac{\mathcal{P}\left(y,\Theta\right)}{\mathcal{P}\left(\Theta\right)}\right)\right\}\\
=& \text{argmax}_{\Theta}\left\{\sum_{y}\mathcal{P}\left(y|\mathcal{X},\Theta_{n}\right)\ln\left(\frac{\mathcal{P}\left(\mathcal{X},y,\Theta\right)}{\mathcal{P}\left(\Theta\right)}\right)\right\}\\
=& \text{argmax}_{\Theta}\left\{\sum_{y}\mathcal{P}\left(y|\mathcal{X},\Theta_{n}\right)\ln\left(\mathcal{P}\left(\mathcal{X},y|\Theta\right)\right)\right\}\\
=& \text{argmax}_{\Theta}\left\{E_{y|\mathcal{X},\Theta_{n}}\left[\ln\left(\mathcal{P}\left(\mathcal{X},y|\Theta\right)\right)\right]\right\}
\end{aligned}
$$

Then $\text{argmax}_{\Theta}\left\{l\left(\Theta|\Theta_{n}\right)\right\} \approx \text{argmax}_{\Theta}\left\{E_{y|\mathcal{X},\Theta_{n}}\left[\ln\left(\mathcal{P}\left(\mathcal{X},y|\Theta\right)\right)\right]\right\}$

Outline

The EM-Algorithm

Steps of EM

1. Expectation under hidden variables.
2. Maximization of the resulting formula.

E-Step

Determine the conditional expectation, $E_{y|\mathcal{X},\Theta_n}\left[\ln\left(\mathcal{P}\left(\mathcal{X},y|\Theta\right)\right)\right]$.

M-Step

Maximize this expression with respect to Θ.

The EM-Algorithm

Steps of EM

1. Expectation under hidden variables.
2. Maximization of the resulting formula.

E-Step

Determine the conditional expectation, $E_{y|\mathcal{X},\Theta_n}\left[\ln\left(\mathcal{P}\left(\mathcal{X},y|\Theta\right)\right)\right]$.

M-Step

Maximize this expression with respect to Θ.

The EM-Algorithm

Steps of EM

1. Expectation under hidden variables.
2. Maximization of the resulting formula.

E-Step

Determine the conditional expectation, $E_{y|\mathcal{X},\Theta_n}\left[\ln\left(\mathcal{P}\left(\mathcal{X},y|\Theta\right)\right)\right]$.

M-Step

Maximize this expression with respect to Θ.

The EM-Algorithm

Steps of EM

1. Expectation under hidden variables.
2. Maximization of the resulting formula.

E-Step

Determine the conditional expectation, $E_{y|\mathcal{X},\Theta_n}\left[\ln\left(\mathcal{P}\left(\mathcal{X},y|\Theta\right)\right)\right]$.

M-Step

Maximize this expression with respect to Θ.

Outline

Cinvestav

Notes and Convergence of EM

Gains between $\mathcal{L}(\Theta)$ and $l(\Theta|\Theta_n)$

Using the hidden variables it is possible to simplify the optimization of $\mathcal{L}(\Theta)$ through $l(\Theta|\Theta_n)$.

Convergence

- Remember that Θ_{n+1} is the estimate for Θ which maximizes the difference $\Delta(\Theta|\Theta_n)$.

Notes and Convergence of EM

Gains between $\mathcal{L}(\Theta)$ and $l(\Theta|\Theta_n)$

Using the hidden variables it is possible to simplify the optimization of $\mathcal{L}(\Theta)$ through $l(\Theta|\Theta_n)$.

Convergence

- Remember that $\Theta_{n+1}$is the estimate for Θ which maximizes the difference $\Delta(\Theta|\Theta_n)$.

Cinvestav

Notes and Convergence of EM

Gains between $\mathcal{L}(\Theta)$ and $l(\Theta|\Theta_n)$

Using the hidden variables it is possible to simplify the optimization of $\mathcal{L}(\Theta)$ through $l(\Theta|\Theta_n)$.

Convergence

- Remember that $\Theta_{n+1}$is the estimate for Θ which maximizes the difference $\Delta(\Theta|\Theta_n)$.

Cinvestav

Therefore

Then, we have

Given the initial estimate of Θ by Θ_n

$$\Delta\left(\Theta_n|\Theta_n\right)=0$$

Now

If we choose Θ_{n+1} to maximize the $\Delta\left(\Theta|\Theta_n\right)$, then

$$\Delta\left(\Theta_{n+1}|\Theta_n\right)\geq\Delta\left(\Theta_n|\Theta_n\right)=0$$

We have that

The Likelihood $\mathcal{L}\left(\Theta\right)$ is not a decreasing function with respect to Θ.

Cinvestav

Therefore

Then, we have

Given the initial estimate of Θ by Θ_n

$$\Delta\left(\Theta_n|\Theta_n\right) = 0$$

Now

If we choose Θ_{n+1} to maximize the $\Delta\left(\Theta|\Theta_n\right)$, then

$$\Delta\left(\Theta_{n+1}|\Theta_n\right) \geq \Delta\left(\Theta_n|\Theta_n\right) = 0$$

We have that

The Likelihood $\mathcal{L}\left(\Theta\right)$ is not a decreasing function with respect to Θ.

Therefore

Then, we have

Given the initial estimate of Θ by Θ_n

$$\Delta\left(\Theta_n|\Theta_n\right) = 0$$

Now

If we choose Θ_{n+1} to maximize the $\Delta\left(\Theta|\Theta_n\right)$, then

$$\Delta\left(\Theta_{n+1}|\Theta_n\right) \geq \Delta\left(\Theta_n|\Theta_n\right) = 0$$

We have that

The Likelihood $\mathcal{L}\left(\Theta\right)$ is not a decreasing function with respect to Θ.

Notes and Convergence of EM

Properties

When the algorithm reaches a fixed point for some Θ_n, the value maximizes $l\left(\Theta|\Theta_n\right)$.

Definition

A fixed point of a function is an element on domain that is mapped to itself by the function:

$$f\left(x\right) = x$$

Basically the EM algorithm does the following

$$EM\left[\Theta^*\right] = \Theta^*$$

Notes and Convergence of EM

Properties

When the algorithm reaches a fixed point for some Θ_n, the value maximizes $l\left(\Theta|\Theta_n\right)$.

Definition

A fixed point of a function is an element on domain that is mapped to itself by the function:

$$f\left(\boldsymbol{x}\right) = \boldsymbol{x}$$

Basically the EM algorithm does the following

$$EM\left[\Theta^*\right] = \Theta^*$$

Notes and Convergence of EM

Properties

When the algorithm reaches a fixed point for some Θ_n, the value maximizes $l\left(\Theta|\Theta_n\right)$.

Definition

A fixed point of a function is an element on domain that is mapped to itself by the function:

$$f\left(\boldsymbol{x}\right) = \boldsymbol{x}$$

Basically the EM algorithm does the following

$$EM\left[\Theta^*\right] = \Theta^*$$

At this moment

We have that

The algorithm reaches a fixed point for some Θ_n, the value Θ^* maximizes $l\left(\Theta|\Theta_n\right)$.

Then, when the algorithm

- It reaches a fixed point for some Θ_n the value maximizes $l\left(\Theta|\Theta_n\right)$.
  - Basically $\Theta_{n+1} = \Theta_n$.

At this moment

We have that

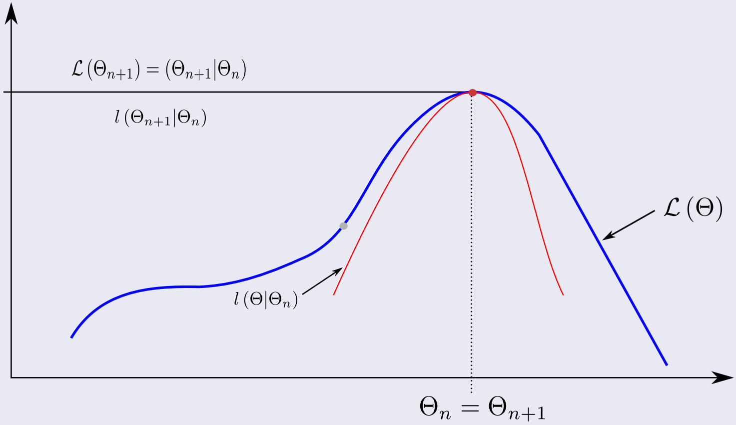
The algorithm reaches a fixed point for some Θ_n, the value Θ^* maximizes $l\left(\Theta|\Theta_n\right)$.

Then, when the algorithm

- It reaches a fixed point for some Θ_n the value maximizes $l\left(\Theta|\Theta_n\right)$.
  - Basically $\Theta_{n+1} = \Theta_n$.

Therefore

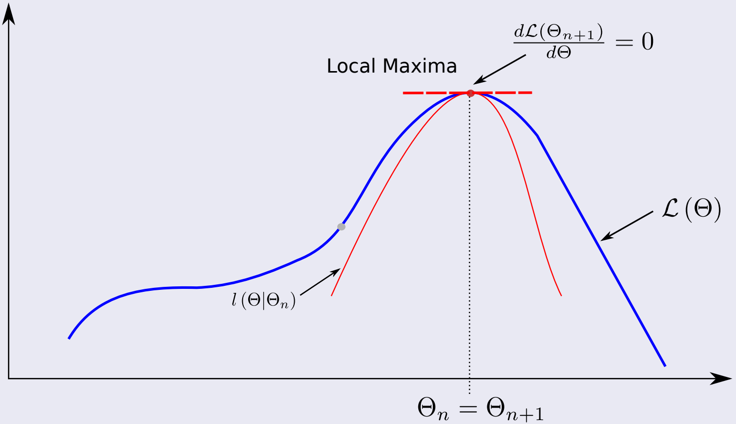
We have



Then

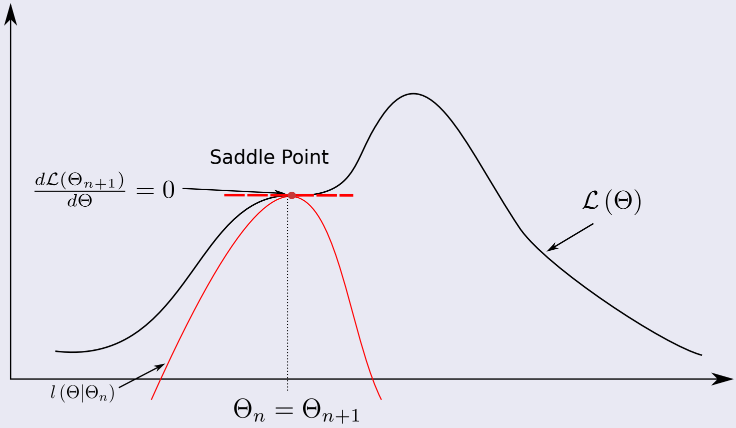
If $\mathcal{L}$ and l are differentiable at Θ_n

- Since $\mathcal{L}$ and l are equal at Θ_n
  - Then, Θ_n is a stationary point of $\mathcal{L}$ i.e. the derivative of $\mathcal{L}$ vanishes at that point.



However

You could finish with the following case, no local maxima



For more on the subject

Please take a look to

Geoffrey McLachlan and Thriyambakam Krishnan, "*The EM Algorithm and Extensions*," John Wiley & Sons, New York, 1996.

Cinvestav

Finding Maximum Likelihood Mixture Densities Parameters via EM

Something Notable

The mixture-density parameter estimation problem is probably one of the most widely used applications of the EM algorithm in the computational pattern recognition community.

We have

$$p(\boldsymbol{x}|\Theta) = \sum_{i=1}^{M} \alpha_i p_i(\boldsymbol{x}|\theta_i) \quad (28)$$

Finding Maximum Likelihood Mixture Densities Parameters via EM

Something Notable

The mixture-density parameter estimation problem is probably one of the most widely used applications of the EM algorithm in the computational pattern recognition community.

We have

$$p\left(\boldsymbol{x}|\Theta\right)=\sum_{i=1}^{M}\alpha_{i}p_{i}\left(\boldsymbol{x}|\theta_{i}\right) \tag{28}$$

where

1. $\Theta=\left(\alpha_{1},...,\alpha_{M},\theta_{1},...,\theta_{M}\right)$
2. $\sum_{i=1}^{M}\alpha_{i}=1$
3. Each p_{i} is a density function parametrized by θ_{i}.

Finding Maximum Likelihood Mixture Densities Parameters via EM

Something Notable

The mixture-density parameter estimation problem is probably one of the most widely used applications of the EM algorithm in the computational pattern recognition community.

We have

$$p(\boldsymbol{x}|\Theta) = \sum_{i=1}^{M} \alpha_i p_i(\boldsymbol{x}|\theta_i) \tag{28}$$

where

1. $\Theta = (\alpha_1, ..., \alpha_M, \theta_1, ..., \theta_M)$
2. $\sum_{i=1}^{M} \alpha_i = 1$
3. Each p_i is a density function parametrized by θ_i.

Finding Maximum Likelihood Mixture Densities Parameters via EM

Something Notable

The mixture-density parameter estimation problem is probably one of the most widely used applications of the EM algorithm in the computational pattern recognition community.

We have

$$p(\boldsymbol{x}|\Theta) = \sum_{i=1}^{M} \alpha_i p_i(\boldsymbol{x}|\theta_i) \tag{28}$$

where

1. $\Theta = (\alpha_1, ..., \alpha_M, \theta_1, ..., \theta_M)$
2. $\sum_{i=1}^{M} \alpha_i = 1$
3. Each p_i is a density function parametrized by θ_i.

A log-likelihood for this function

We have

$$\log \mathcal{L}\left(\Theta|\mathcal{X}\right) = \log \prod_{i=1}^{N} p\left(x_i|\Theta\right) = \sum_{i=1}^{N} \log \left(\sum_{j=1}^{M} \alpha_j p_j \left(x_i|\theta_j\right) \right) \quad (29)$$

Note: This is too difficult to optimize due to the log function.

However

We can simplify this assuming the following:

- We assume that each unobserved data $\mathcal{Y} = \{y_i\}_{i=1}^{N}$ has a the following range $y_i \in \{1, ..., M\}$
- $y_i = k$ if the i^{th} samples was generated by the k^{th} mixture.

Cinvestav

A log-likelihood for this function

We have

$$\log \mathcal{L}\left(\Theta|\mathcal{X}\right) = \log \prod_{i=1}^{N} p\left(x_i|\Theta\right) = \sum_{i=1}^{N} \log\left(\sum_{j=1}^{M} \alpha_j p_j\left(x_i|\theta_j\right)\right) \tag{29}$$

Note: This is too difficult to optimize due to the log function.

However

We can simplify this assuming the following:

1. We assume that each unobserved data $\mathcal{Y}=\{y_i\}_{i=1}^{N}$ has a the following range $y_i \in \{1,...,M\}$
2. $y_i = k$ if the i^{th} samples was generated by the k^{th} mixture.

A log-likelihood for this function

We have

$$\log \mathcal{L}\left(\Theta|\mathcal{X}\right) = \log \prod_{i=1}^{N} p\left(x_i|\Theta\right) = \sum_{i=1}^{N} \log\left(\sum_{j=1}^{M} \alpha_j p_j\left(x_i|\theta_j\right)\right) \quad (29)$$

Note: This is too difficult to optimize due to the log function.

However

We can simplify this assuming the following:

1. We assume that each unobserved data $\mathcal{Y} = \{y_i\}_{i=1}^{N}$ has a the following range $y_i \in \{1, ..., M\}$
2. $y_i = k$ if the i^{th} samples was generated by the k^{th} mixture.

A log-likelihood for this function

We have

$$\log \mathcal{L}(\Theta|\mathcal{X}) = \log \prod_{i=1}^{N} p(x_i|\Theta) = \sum_{i=1}^{N} \log \left(\sum_{j=1}^{M} \alpha_j p_j (x_i|\theta_j) \right) \quad (29)$$

Note: This is too difficult to optimize due to the log function.

However

We can simplify this assuming the following:

1. We assume that each unobserved data $\mathcal{Y} = \{y_i\}_{i=1}^{N}$ has a the following range $y_i \in \{1, ..., M\}$
2. $y_i = k$ if the i^{th} samples was generated by the k^{th} mixture.

Cinvestav

A log-likelihood for this function

We have

$$\log \mathcal{L}(\Theta|\mathcal{X}) = \log \prod_{i=1}^{N} p(x_i|\Theta) = \sum_{i=1}^{N} \log \left(\sum_{j=1}^{M} \alpha_j p_j (x_i|\theta_j) \right) \quad (29)$$

Note: This is too difficult to optimize due to the log function.

However

We can simplify this assuming the following:

1. We assume that each unobserved data $\mathcal{Y} = \{y_i\}_{i=1}^{N}$ has a the following range $y_i \in \{1, ..., M\}$
2. $y_i = k$ if the i^{th} samples was generated by the k^{th} mixture.

Cinvestav

Now

We have

$$\log \mathcal{L}\left(\Theta|\mathcal{X},\mathcal{Y}\right) = \log\left[P\left(\mathcal{X},\mathcal{Y}|\Theta\right)\right] \tag{30}$$

Remember that $\mathcal{X} = \{x_1, x_2, ..., x_N\}$ with $\mathcal{Y} = \{y_1, y_2, ..., y_N\}$ and assuming independence

$$\begin{aligned}
\log\left[P\left(\mathcal{X},\mathcal{Y}|\Theta\right)\right] &= \log\left[P\left(x_1, x_2, \dots, x_N, y_1, y_2, \dots, y_N|\Theta\right)\right]\\
&= \log\left[P\left(x_1, y_1, \dots, x_i, y_i, \dots, x_N, y_N|\Theta\right)\right]\\
&= \log\prod_{i=1}^{N} P\left(x_i, y_i|\Theta\right)\\
&= \sum_{i=1}^{N} \log P\left(x_i, y_i|\Theta\right)
\end{aligned}$$

Now

We have

$$\log \mathcal{L}(\Theta|\mathcal{X},\mathcal{Y}) = \log\left[P(\mathcal{X},\mathcal{Y}|\Theta)\right] \tag{30}$$

Remember that $\mathcal{X} = \{x_1, x_2, ..., x_N\}$ with $\mathcal{Y} = \{y_1, y_2, ..., y_N\}$ and assuming independence

$$\begin{aligned}
\log\left[P(\mathcal{X},\mathcal{Y}|\Theta)\right] &= \log\left[P(x_1, x_2, ..., x_N, y_1, y_2, ..., y_N|\Theta)\right] \\
&= \log\left[P(x_1, y_1, ..., x_i, y_i, ..., x_N, y_N|\Theta)\right] \\
&= \log\prod_{i=1}^{N} P(x_i, y_i|\Theta) \\
&= \sum_{i=1}^{N} \log P(x_i, y_i|\Theta)
\end{aligned}$$

Now

We have

$$\log \mathcal{L}\left(\Theta|\mathcal{X},\mathcal{Y}\right) = \log\left[P\left(\mathcal{X},\mathcal{Y}|\Theta\right)\right] \tag{30}$$

Remember that $\mathcal{X} = \{x_1, x_2, ..., x_N\}$ with $\mathcal{Y} = \{y_1, y_2, ..., y_N\}$ and assuming independence

$$\begin{aligned}
\log\left[P\left(\mathcal{X},\mathcal{Y}|\Theta\right)\right] &= \log\left[P\left(x_1, x_2, ..., x_N, y_1, y_2, ..., y_N|\Theta\right)\right] \\
&= \log\left[P\left(x_1, y_1, ..., x_i, y_i, ..., x_N, y_N|\Theta\right)\right] \\
&= \log\prod_{i=1}^{N} P\left(x_i, y_i|\Theta\right) \\
&= \sum_{i=1}^{N} \log P\left(x_i, y_i|\Theta\right)
\end{aligned}$$

Now

We have

$$\log \mathcal{L}(\Theta|\mathcal{X},\mathcal{Y}) = \log\left[P(\mathcal{X},\mathcal{Y}|\Theta)\right] \tag{30}$$

Remember that $\mathcal{X} = \{x_1, x_2, ..., x_N\}$ with $\mathcal{Y} = \{y_1, y_2, ..., y_N\}$ and assuming independence

$$\begin{aligned}
\log\left[P(\mathcal{X},\mathcal{Y}|\Theta)\right] &= \log\left[P(x_1, x_2, ..., x_N, y_1, y_2, ..., y_N|\Theta)\right] \\
&= \log\left[P(x_1, y_1, ..., x_i, y_i, ..., x_N, y_N|\Theta)\right] \\
&= \log \prod_{i=1}^{N} P(x_i, y_i|\Theta) \\
&= \sum_{i=1}^{N} \log P(x_i, y_i|\Theta)
\end{aligned}$$

Now

We have

$$\log \mathcal{L}\left(\Theta|\mathcal{X},\mathcal{Y}\right)=\log\left[P\left(\mathcal{X},\mathcal{Y}|\Theta\right)\right] \tag{30}$$

Remember that $\mathcal{X}=\{x_1,x_2,...,x_N\}$ with $\mathcal{Y}=\{y_1,y_2,...,y_N\}$ and assuming independence

$$\begin{aligned}\log\left[P\left(\mathcal{X},\mathcal{Y}|\Theta\right)\right] &= \log\left[P\left(x_1,x_2,...,x_N,y_1,y_2,...,y_N|\Theta\right)\right]\\ &= \log\left[P\left(x_1,y_1,...,x_i,y_i,...,x_N,y_N|\Theta\right)\right]\\ &= \log\prod_{i=1}^{N}P\left(x_i,y_i|\Theta\right)\\ &= \sum_{i=1}^{N}\log P\left(x_i,y_i|\Theta\right)\end{aligned}$$

Then

Thus, by the chain Rule

$$\sum_{i=1}^{N} \log P\left(x_i, y_i|\Theta\right) = \sum_{i=1}^{N} \log\left[P\left(x_i|y_i, \theta_{y_i}\right) P\left(y_i|\theta_{y_i}\right)\right] \tag{31}$$

Question Do you need y_i if you know θ_{y_i} or the other way around?

Finally

$$\sum_{i=1}^{N} \log\left[P\left(x_i|y_i, \theta_{y_i}\right) P\left(y_i|\theta_{y_i}\right)\right] = \sum_{i=1}^{N} \log\left[P\left(y_i\right) p_{y_i}\left(x_i|\theta_{y_i}\right)\right] \tag{32}$$

NOPE: You do not need y_i if you know θ_{y_i} or the other way around.

Then

Thus, by the chain Rule

$$\sum_{i=1}^{N} \log P\left(x_i, y_i|\Theta\right) = \sum_{i=1}^{N} \log\left[P\left(x_i|y_i, \theta_{y_i}\right) P\left(y_i|\theta_{y_i}\right)\right] \tag{31}$$

Question Do you need y_i if you know θ_{y_i} or the other way around?

Finally

$$\sum_{i=1}^{N} \log\left[P\left(x_i|y_i, \theta_{y_i}\right) P\left(y_i|\theta_{y_i}\right)\right] = \sum_{i=1}^{N} \log\left[P\left(y_i\right) p_{y_i}\left(x_i|\theta_{y_i}\right)\right] \tag{32}$$

NOPE: You do not need y_i if you know θ_{y_i} or the other way around.

Cinvestav

Finally, we have

Making $\alpha_{y_i} = P(y_i)$

$$\log \mathcal{L}(\Theta|\mathcal{X}, \mathcal{Y}) = \sum_{i=1}^{N} \log \left[\alpha_{y_i} P(x_i|y_i, \theta_{y_i})\right] \quad (33)$$

Cinvestav

Problem

Which Labels?

We do not know the values of $\mathcal{Y}$.

We can get away by using the following idea

Assume the $\mathcal{Y}$ is a random variable.

Problem

Which Labels?

We do not know the values of $\mathcal{Y}$.

We can get away by using the following idea

Assume the $\mathcal{Y}$ is a random variable.

Outline

Cinvestav

Thus

You do a first guess for the parameters at the beginning of EM

$$\Theta^g = (\alpha_1^g, ..., \alpha_M^g, \theta_1^g, ..., \theta_M^g) \tag{34}$$

Then, it is possible to calculate given the parametric probability

$$p_j\left(x_i|\theta_j^g\right)$$

Therefore

The mixing parameters α_j can be though of as a prior probabilities of each mixture:

$$\alpha_j = p\left(\text{component } j\right) \tag{35}$$

Thus

You do a first guess for the parameters at the beginning of EM

$$\Theta^g = \left(\alpha_1^g, ..., \alpha_M^g, \theta_1^g, ..., \theta_M^g\right) \tag{34}$$

Then, it is possible to calculate given the parametric probability

$$p_j\left(x_i|\theta_j^g\right)$$

Therefore

The mixing parameters α_j can be though of as a prior probabilities of each mixture:

$$\alpha_j = p\left(\text{component } j\right) \tag{35}$$

Thus

You do a first guess for the parameters at the beginning of EM

$$\Theta^g = (\alpha_1^g, ..., \alpha_M^g, \theta_1^g, ..., \theta_M^g) \quad (34)$$

Then, it is possible to calculate given the parametric probability

$$p_j\left(x_i|\theta_j^g\right)$$

Therefore

The mixing parameters α_j can be though of as a prior probabilities of each mixture:

$$\alpha_j = p\left(\text{component } j\right) \quad (35)$$

Cinvestav

Outline

Outline

Cinvestav

We want to calculate the following probability

We want to calculate

$$p\left(y_i|x_i, \Theta^g\right)$$

Basically

We want a Bayesian formulation of this probability.

- Assuming that the $\boldsymbol{y} = \left(y_1, y_2, ..., y_N\right)$ are samples identically independent samples from a distribution.

We want to calculate the following probability

We want to calculate

$$p\left(y_i|x_i,\Theta^g\right)$$

Basically

We want a Bayesian formulation of this probability.

- Assuming that the $y=(y_1,y_2,...,y_N)$ are samples identically independent samples from a distribution.

We want to calculate the following probability

We want to calculate

$$p\left(y_i|x_i,\Theta^g\right)$$

Basically

We want a Bayesian formulation of this probability.

- Assuming that the $\boldsymbol{y} = \left(y_1, y_2, ..., y_N\right)$ are samples identically independent samples from a distribution.

Using Bayes' Rule

Compute

$$
\begin{aligned}
p\left(y_i|x_i,\Theta^g\right) &= \frac{p\left(y_i,x_i|\Theta^g\right)}{p\left(x_i|\Theta^g\right)}\\
&= \frac{p\left(x_i|\Theta^g\right)p\left(y_i|\theta_{y_i}^g\right)}{p\left(x_i|\Theta^g\right)} \quad \text{We know } \theta_{y_i}^g \Rightarrow \text{Drop it}\\
&= \frac{\alpha_{y_i}^g p_{y_i}\left(x_i|\theta_{y_i}^g\right)}{p\left(x_i|\Theta^g\right)}\\
&= \frac{\alpha_{y_i}^g p_{y_i}\left(x_i|\theta_{y_i}^g\right)}{\sum_{k=1}^M \alpha_k^g p_k\left(x_i|\theta_k^g\right)}
\end{aligned}
$$

Cinvestav

Using Bayes' Rule

Compute

$$
\begin{aligned}
p\left(y_i \middle| x_i, \Theta^g\right) &= \frac{p\left(y_i, x_i \middle| \Theta^g\right)}{p\left(x_i \middle| \Theta^g\right)} \\
&= \frac{p\left(x_i \middle| \Theta^g\right) p\left(y_i \middle| \theta^g_{y_i}\right)}{p\left(x_i \middle| \Theta^g\right)} \text{ We know } \theta^g_{y_i} \Rightarrow \text{Drop it} \\
&= \frac{\alpha^g_{y_i} p_{y_i}\left(x_i \middle| \theta^g_{y_i}\right)}{p\left(x_i \middle| \Theta^g\right)} \\
&= \frac{\alpha^g_{y_i} p_{y_i}\left(x_i \middle| \theta^g_{y_i}\right)}{\sum_{k=1}^{M} \alpha^g_k p_k\left(x_i \middle| \theta^g_k\right)}
\end{aligned}
$$

Cinvestav

Using Bayes' Rule

Compute

$$
\begin{aligned}
p\left(y_i | x_i, \Theta^g\right) &= \frac{p\left(y_i, x_i | \Theta^g\right)}{p\left(x_i | \Theta^g\right)} \\
&= \frac{p\left(x_i | \Theta^g\right) p\left(y_i | \theta^g_{y_i}\right)}{p\left(x_i | \Theta^g\right)} \text{ We know } \theta^g_{y_i} \Rightarrow \text{ Drop it} \\
&= \frac{\alpha^g_{y_i} p_{y_i}\left(x_i | \theta^g_{y_i}\right)}{p\left(x_i | \Theta^g\right)} \\
&= \frac{\alpha^g_{y_i} p_{y_i}\left(x_i | \theta^g_{y_i}\right)}{\sum_{k=1}^{M} \alpha^g_k p_k\left(x_i | \theta^g_k\right)}
\end{aligned}
$$

Using Bayes' Rule

Compute

$$
\begin{aligned}
p\left(y_i | x_i, \Theta^g\right) &= \frac{p\left(y_i, x_i | \Theta^g\right)}{p\left(x_i | \Theta^g\right)} \\
&= \frac{p\left(x_i | \Theta^g\right) p\left(y_i | \theta^g_{y_i}\right)}{p\left(x_i | \Theta^g\right)} \text{ We know } \theta^g_{y_i} \Rightarrow \text{ Drop it} \\
&= \frac{\alpha^g_{y_i} p_{y_i}\left(x_i | \theta^g_{y_i}\right)}{p\left(x_i | \Theta^g\right)} \\
&= \frac{\alpha^g_{y_i} p_{y_i}\left(x_i | \theta^g_{y_i}\right)}{\sum_{k=1}^{M} \alpha^g_k p_k\left(x_i | \theta^g_k\right)}
\end{aligned}
$$

As in Naive Bayes

We have the fact that there is a probability per probability at the mixture and sample

$$p\left(y_i|x_i,\Theta^g\right)=\frac{\alpha_{y_i}^g p_{y_i}\left(x_i|\theta_{y_i}^g\right)}{\sum_{k=1}^M \alpha_k^g p_k\left(x_i|\theta_k^g\right)}\ \forall x_i,\ y_i \text{ and } k\in\{1,...,M\}$$

This is going to be updated at each iteration of the EM algorithm

After the initial Guess!!! Until convergence!!!

As in Naive Bayes

We have the fact that there is a probability per probability at the mixture and sample

$$p\left(y_i|x_i,\Theta^g\right)=\frac{\alpha^g_{y_i}p_{y_i}\left(x_i|\theta^g_{y_i}\right)}{\sum_{k=1}^{M}\alpha^g_k p_k\left(x_i|\theta^g_k\right)}\ \forall x_i,\ y_i \text{ and } k\in\{1,...,M\}$$

This is going to be updated at each iteration of the EM algorithm

After the initial Guess!!! Until convergence!!!

Additionally

We assume again that the samples $y_i's$ are identically and independent samples

$$p\left(\boldsymbol{y}|\mathcal{X},\Theta^g\right)=\prod_{i=1}^{N}p\left(y_i|x_i,\Theta^g\right) \tag{36}$$

Where $\boldsymbol{y}=\left(y_1,y_2,...,y_N\right)$

Now, using equation 17

Then

$$Q\left(\Theta|\Theta^g\right)=\sum_{\boldsymbol{y}\in\mathcal{Y}}\log\left(\mathcal{L}\left(\Theta|\mathcal{X},\boldsymbol{y}\right)\right)p\left(\boldsymbol{y}|\mathcal{X},\Theta^g\right)$$

$$=\sum_{\boldsymbol{y}\in\mathcal{Y}}\sum_{i=1}^{N}\log\left[\alpha_{y_i}p_{y_i}\left(x_i|\theta_{y_i}\right)\right]\prod_{j=1}^{N}p\left(y_j|x_j,\Theta^g\right)$$

Now, using equation 17

Then

$$Q\left(\Theta|\Theta^{g}\right) = \sum_{\boldsymbol{y}\in\mathcal{Y}} \log\left(\mathcal{L}\left(\Theta|\mathcal{X},\boldsymbol{y}\right)\right) p\left(\boldsymbol{y}|\mathcal{X},\Theta^{g}\right)$$

$$= \sum_{\boldsymbol{y}\in\mathcal{Y}} \sum_{i=1}^{N} \log\left[\alpha_{y_i} p_{y_i}\left(x_i|\theta_{y_i}\right)\right] \prod_{j=1}^{N} p\left(y_j|x_j,\Theta^{g}\right)$$

Here, a small stop

What is the meaning of $\sum_{\boldsymbol{y}\in\mathcal{Y}}$

It is actually a summation of all possible states of the random vector $\boldsymbol{y}$.

Then, we can rewrite the previous summation as

$$\sum_{y\in\mathcal{Y}} = \underbrace{\sum_{y_1=1}^{M}\sum_{y_2=1}^{M}\cdots\sum_{y_N=1}^{M}}_{N}$$

Running over all the samples $\{x_1, x_2, ..., x_N\}$.

Here, a small stop

What is the meaning of $\sum_{\boldsymbol{y}\in\mathcal{Y}}$

It is actually a summation of all possible states of the random vector $\boldsymbol{y}$.

Then, we can rewrite the previous summation as

$$\sum_{\boldsymbol{y}\in\mathcal{Y}} = \underbrace{\sum_{y_1=1}^{M}\sum_{y_2=1}^{M}\cdots\sum_{y_N=1}^{M}}_{N}$$

Running over all the samples $\{x_1, x_2, ..., x_N\}$.

Then

We have

$$Q\left(\Theta|\Theta^{g}\right)=\sum_{y_1=1}^{M}\sum_{y_2=1}^{M}\cdots\sum_{y_N=1}^{M}\sum_{i=1}^{N}\left[\log\left[\alpha_{y_i}p_{y_i}\left(x_i|\theta_{y_i}\right)\right]\prod_{j=1}^{N}p\left(y_j|x_j,\Theta^{g}\right)\right]$$

We introduce the following

We have the following function

$$\delta_{l,y_i} = \begin{cases} 1 & I = y_i \\ 0 & I \neq y_i \end{cases}$$

Therefore, we can do the following

$$\alpha_i = \sum_{j=1}^{M} \delta_{i,j} \alpha_j$$

Then

$$\log\left[\alpha_{y_i} p_{y_i}\left(x_i|\theta_{y_i}\right)\right] \prod_{j=1}^{N} p\left(y_j|x_j, \Theta^g\right) = \sum_{l=1}^{M} \delta_{l,y_i} \log\left[\alpha_l p_l\left(x_i|\theta_l\right)\right] \prod_{j=1}^{N} p\left(y_j|x_j, \Theta^g\right)$$

We introduce the following

We have the following function

$$\delta_{l,y_i} = \begin{cases} 1 & I = y_i \\ 0 & I \neq y_i \end{cases}$$

Therefore, we can do the following

$$\alpha_i = \sum_{j=1}^{M} \delta_{i,j} \alpha_j$$

Then

$$\log\left[\alpha_{y_i} p_{y_i}\left(x_i|\theta_{y_i}\right)\right] \prod_{j=1}^{N} p\left(y_j|x_j, \Theta^g\right) = \sum_{l=1}^{M} \delta_{l,y_i} \log\left[\alpha_l p_l\left(x_i|\theta_l\right)\right] \prod_{j=1}^{N} p\left(y_j|x_j, \Theta^g\right)$$

We introduce the following

We have the following function

$$\delta_{l,y_i} = \begin{cases} 1 & I = y_i \\ 0 & I \neq y_i \end{cases}$$

Therefore, we can do the following

$$\alpha_i = \sum_{j=1}^{M} \delta_{i,j} \alpha_j$$

Then

$$\log\left[\alpha_{y_i} p_{y_i}\left(x_i|\theta_{y_i}\right)\right] \prod_{j=1}^{N} p\left(y_j|x_j, \Theta^g\right) = \sum_{l=1}^{M} \delta_{l,y_i} \log\left[\alpha_l p_l\left(x_i|\theta_l\right)\right] \prod_{j=1}^{N} p\left(y_j|x_j, \Theta^g\right)$$

Thus

We have that for
$\sum_{y_1=1}^{M} \cdots \sum_{y_N=1}^{M} \sum_{i=1}^{N} \log\left[\alpha_{y_i} p_{y_i}\left(x_i|\theta_{y_i}\right)\right] \prod_{j=1}^{N} p\left(y_j|x_j,\Theta^g\right) = *$

$$* = \sum_{y_1=1}^{M} \sum_{y_2=1}^{M} \cdots \sum_{y_N=1}^{M} \sum_{i=1}^{N} \sum_{l=1}^{M} \delta_{l,y_i} \log\left[\alpha_l p_l\left(x_i|\theta_l\right)\right] \prod_{j=1}^{N} p\left(y_j|x_j,\Theta^g\right)$$

$$= \sum_{i=1}^{N} \sum_{l=1}^{M} \log\left[\alpha_l p_l\left(x_i|\theta_l\right)\right] \sum_{y_1=1}^{M} \sum_{y_2=1}^{M} \cdots \sum_{y_N=1}^{M} \left[\delta_{l,y_i} \prod_{j=1}^{N} p\left(y_j|x_j,\Theta^g\right)\right]$$

Because

$\sum_{y_1=1}^{M} \sum_{y_2=1}^{M} \cdots \sum_{y_N=1}^{M}$ applies only to $\delta_{l,y_i} \prod_{j=1}^{N} p\left(y_j|x_j,\Theta^g\right)$

Thus

We have that for
$\sum_{y_1=1}^{M} \cdots \sum_{y_N=1}^{M} \sum_{i=1}^{N} \log\left[\alpha_{y_i} p_{y_i}\left(x_i|\theta_{y_i}\right)\right] \prod_{j=1}^{N} p\left(y_j|x_j,\Theta^g\right) = *$

$$* = \sum_{y_1=1}^{M}\sum_{y_2=1}^{M}\cdots\sum_{y_N=1}^{M}\sum_{i=1}^{N}\sum_{l=1}^{M}\delta_{l,y_i}\log\left[\alpha_l p_l\left(x_i|\theta_l\right)\right]\prod_{j=1}^{N}p\left(y_j|x_j,\Theta^g\right)$$
$$= \sum_{i=1}^{N}\sum_{l=1}^{M}\log\left[\alpha_l p_l\left(x_i|\theta_l\right)\right]\sum_{y_1=1}^{M}\sum_{y_2=1}^{M}\cdots\sum_{y_N=1}^{M}\left[\delta_{l,y_i}\prod_{j=1}^{N}p\left(y_j|x_j,\Theta^g\right)\right]$$

Because

$\sum_{y_1=1}^{M}\sum_{y_2=1}^{M}\cdots\sum_{y_N=1}^{M}$ applies only to $\delta_{l,y_i}\prod_{j=1}^{N}p\left(y_j|x_j,\Theta^g\right)$

Thus

We have that for
$\sum_{y_1=1}^{M} \cdots \sum_{y_N=1}^{M} \sum_{i=1}^{N} \log\left[\alpha_{y_i} p_{y_i}\left(x_i|\theta_{y_i}\right)\right] \prod_{j=1}^{N} p\left(y_j|x_j, \Theta^g\right) = *$

$$* = \sum_{y_1=1}^{M} \sum_{y_2=1}^{M} \cdots \sum_{y_N=1}^{M} \sum_{i=1}^{N} \sum_{l=1}^{M} \delta_{l,y_i} \log\left[\alpha_l p_l\left(x_i|\theta_l\right)\right] \prod_{j=1}^{N} p\left(y_j|x_j, \Theta^g\right)$$

$$= \sum_{i=1}^{N} \sum_{l=1}^{M} \log\left[\alpha_l p_l\left(x_i|\theta_l\right)\right] \sum_{y_1=1}^{M} \sum_{y_2=1}^{M} \cdots \sum_{y_N=1}^{M} \left[\delta_{l,y_i} \prod_{j=1}^{N} p\left(y_j|x_j, \Theta^g\right)\right]$$

Because

$\sum_{y_1=1}^{M} \sum_{y_2=1}^{M} \cdots \sum_{y_N=1}^{M}$ applies only to $\delta_{l,y_i} \prod_{j=1}^{N} p\left(y_j|x_j, \Theta^g\right)$

Cinvestav

Then, we have that

First notice the following

$$\sum_{y_1=1}^{M}\sum_{y_2=1}^{M}\cdots\sum_{y_N=1}^{M}\left[\delta_{l,y_i}\prod_{j=1}^{N}p\left(y_j|x_j,\Theta^g\right)\right]=$$

$$=\left(\sum_{y_1=1}^{M}\cdots\sum_{y_{i-1}=1}^{M}\sum_{y_{i+1}=1}^{M}\cdots\sum_{y_N=1}^{M}\left\{\left[\sum_{y_i=1}^{M}\delta_{l,y_i}p\left(y_i|x_i,\Theta^g\right)\right]\prod_{j=1,j\neq i}^{N}p\left(y_j|x_j,\Theta^g\right)\right\}\right)$$

Then, we have

$$\sum_{y_i=1}^{M}\delta_{l,y_i}p\left(y_i|x_i,\Theta^g\right)=p\left(l|x_i,\Theta^g\right)$$

Then, we have that

First notice the following

$$\sum_{y_1=1}^{M}\sum_{y_2=1}^{M}\cdots\sum_{y_N=1}^{M}\left[\delta_{l,y_i}\prod_{j=1}^{N}p\left(y_j|x_j,\Theta^g\right)\right]=$$

$$=\left(\sum_{y_1=1}^{M}\cdots\sum_{y_{i-1}=1}^{M}\sum_{y_{i+1}=1}^{M}\cdots\sum_{y_N=1}^{M}\left\{\left[\sum_{y_i=1}^{M}\delta_{l,y_i}p\left(y_i|x_i,\Theta^g\right)\right]\prod_{j=1,j\neq i,}^{N}p\left(y_j|x_j,\Theta^g\right)\right\}\right)$$

Then, we have

$$\sum_{y_i=1}^{M}\delta_{l,y_i}p\left(y_i|x_i,\Theta^g\right)=p\left(l|x_i,\Theta^g\right)$$

Then, we have that

First notice the following

$$\sum_{y_1=1}^{M}\sum_{y_2=1}^{M}\cdots\sum_{y_N=1}^{M}\left[\delta_{l,y_i}\prod_{j=1}^{N}p\left(y_j|x_j,\Theta^g\right)\right]=$$

$$=\left(\sum_{y_1=1}^{M}\cdots\sum_{y_{i-1}=1}^{M}\sum_{y_{i+1}=1}^{M}\cdots\sum_{y_N=1}^{M}\left\{\left[\sum_{y_i=1}^{M}\delta_{l,y_i}p\left(y_i|x_i,\Theta^g\right)\right]\prod_{j=1,j\neq i,}^{N}p\left(y_j|x_j,\Theta^g\right)\right\}\right)$$

Then, we have

$$\sum_{y_i=1}^{M}\delta_{l,y_i}p\left(y_i|x_i,\Theta^g\right)=p\left(l|x_i,\Theta^g\right)$$

In this way

Plugging back the previous equation

$$\sum_{y_1=1}^{M}\sum_{y_2=1}^{M}\cdots\sum_{y_N=1}^{M}\delta_{l,y_i}\prod_{j=1}^{N}p\left(y_j|x_j,\Theta^g\right)=$$

$$=\left(\sum_{y_1=1}^{M}\cdots\sum_{y_{i-1}=1}^{M}\sum_{y_{i+1}=1}^{M}\cdots\sum_{y_N=1}^{M}p\left(l|x_i,\Theta^g\right)\prod_{j=1,j\neq i}^{N}p\left(y_j|x_j,\Theta^g\right)\right)$$

$$=\left(\sum_{y_1=1}^{M}\cdots\sum_{y_{i-1}=1}^{M}\sum_{y_{i+1}=1}^{M}\cdots\sum_{y_N=1}^{M}\prod_{j=1,j\neq i}^{N}p\left(y_j|x_j,\Theta^g\right)\right)p\left(l|x_i,\Theta^g\right)$$

In this way

Plugging back the previous equation

$$\sum_{y_1=1}^{M}\sum_{y_2=1}^{M}\cdots\sum_{y_N=1}^{M}\delta_{l,y_i}\prod_{j=1}^{N}p\left(y_j|x_j,\Theta^g\right)=$$

$$=\left(\sum_{y_1=1}^{M}\cdots\sum_{y_{i-1}=1}^{M}\sum_{y_{i+1}=1}^{M}\cdots\sum_{y_N=1}^{M}p\left(l|x_i,\Theta^g\right)\prod_{j=1,j\neq i}^{N}p\left(y_j|x_j,\Theta^g\right)\right)$$

$$=\left(\sum_{y_1=1}^{M}\cdots\sum_{y_{i-1}=1}^{M}\sum_{y_{i+1}=1}^{M}\cdots\sum_{y_N=1}^{M}\prod_{j=1,j\neq i}^{N}p\left(y_j|x_j,\Theta^g\right)\right)p\left(l|x_i,\Theta^g\right)$$

Cinvestav

In this way

Plugging back the previous equation

$$\sum_{y_1=1}^{M}\sum_{y_2=1}^{M}\cdots\sum_{y_N=1}^{M}\delta_{l,y_i}\prod_{j=1}^{N}p\left(y_j|x_j,\Theta^g\right)=$$

$$=\left(\sum_{y_1=1}^{M}\cdots\sum_{y_{i-1}=1}^{M}\sum_{y_{i+1}=1}^{M}\cdots\sum_{y_N=1}^{M}p\left(l|x_i,\Theta^g\right)\prod_{j=1,j\neq i}^{N}p\left(y_j|x_j,\Theta^g\right)\right)$$

$$=\left(\sum_{y_1=1}^{M}\cdots\sum_{y_{i-1}=1}^{M}\sum_{y_{i+1}=1}^{M}\cdots\sum_{y_N=1}^{M}\prod_{j=1,j\neq i}^{N}p\left(y_j|x_j,\Theta^g\right)\right)p\left(l|x_i,\Theta^g\right)$$

Now, what about...?

The left part of the equation

$$\sum_{y_1=1}^{M}\cdots\sum_{y_{i-1}=1}^{M}\sum_{y_{i+1}=1}^{M}\cdots\sum_{y_N=1}^{M}\prod_{j=1,j\neq i}^{N} p\left(y_j|x_j,\Theta^g\right) =$$

$$= \left[\sum_{y_1=1}^{M} p\left(y_1|x_1,\Theta^g\right)\right]\cdots\left[\sum_{y_{i-1}=1}^{M} p\left(y_{i-1}|x_{i-1},\Theta^g\right)\right]\times\cdots$$

$$\left[\sum_{y_{i+1}=1}^{M} p\left(y_{i+1}|x_{i+1},\Theta^g\right)\right]\cdots\left[\sum_{y_N=1}^{M} p\left(y_N|x_N,\Theta^g\right)\right]$$

$$= \prod_{j=1,j\neq i}^{N}\left[\sum_{y_j=1}^{M} p\left(y_j|x_j,\Theta^g\right)\right]$$

Cinvestav

Now, what about...?

The left part of the equation

$$\sum_{y_1=1}^{M}\cdots\sum_{y_{i-1}=1}^{M}\sum_{y_{i+1}=1}^{M}\cdots\sum_{y_N=1}^{M}\prod_{j=1,j\neq i}^{N} p\left(y_j|x_j,\Theta^g\right) =$$

$$= \left[\sum_{y_1=1}^{M} p\left(y_1|x_1,\Theta^g\right)\right]\cdots\left[\sum_{y_{i-1}=1}^{M} p\left(y_{i-1}|x_{i-1},\Theta^g\right)\right]\times \dots$$

$$\left[\sum_{y_{i+1}=1}^{M} p\left(y_{i+1}|x_{i+1},\Theta^g\right)\right]\cdots\left[\sum_{y_N=1}^{M} p\left(y_N|x_N,\Theta^g\right)\right]$$

$$= \prod_{j=1,j\neq i}^{N}\left[\sum_{y_j=1}^{M} p\left(y_j|x_j,\Theta^g\right)\right]$$

Now, what about...?

The left part of the equation

$$
\begin{aligned}
&\sum_{y_1=1}^{M} \cdots \sum_{y_{i-1}=1}^{M} \sum_{y_{i+1}=1}^{M} \cdots \sum_{y_N=1}^{M} \prod_{j=1, j\neq i}^{N} p\left(y_j | x_j, \Theta^g\right) = \\
&= \left[\sum_{y_1=1}^{M} p\left(y_1 | x_1, \Theta^g\right)\right] \cdots \left[\sum_{y_{i-1}=1}^{M} p\left(y_{i-1} | x_{i-1}, \Theta^g\right)\right] \times \ldots \\
&\quad \left[\sum_{y_{i+1}=1}^{M} p\left(y_{i+1} | x_{i+1}, \Theta^g\right)\right] \cdots \left[\sum_{y_N=1}^{M} p\left(y_N | x_N, \Theta^g\right)\right] \\
&= \prod_{j=1, j\neq i}^{N} \left[\sum_{y_j=1}^{M} p\left(y_j | x_j, \Theta^g\right)\right]
\end{aligned}
$$

Then, we have that

Plugging back to the original equation

$$\left\{ \sum_{y_1=1}^{M} \cdots \sum_{y_{i-1}=1}^{M} \sum_{y_{i+1}=1}^{M} \cdots \sum_{y_N=1}^{M} \prod_{j=1, j\neq i}^{N} p\left(y_j|x_j, \Theta^g\right) \right\} p\left(l|x_i, \Theta^g\right) =$$

$$= \left\{ \prod_{j=1, j\neq i}^{N} \left[\sum_{y_j=1}^{M} p\left(y_j|x_j, \Theta^g\right) \right] \right\} p\left(l|x_i, \Theta^g\right)$$

Then, we have that

Plugging back to the original equation

$$\left\{ \sum_{y_1=1}^{M} \cdots \sum_{y_{i-1}=1}^{M} \sum_{y_{i+1}=1}^{M} \cdots \sum_{y_N=1}^{M} \prod_{j=1, j\neq i}^{N} p\left(y_j | x_j, \Theta^g\right) \right\} p\left(l | x_i, \Theta^g\right) =$$

$$= \left\{ \prod_{j=1, j\neq i}^{N} \left[\sum_{y_j=1}^{M} p\left(y_j | x_j, \Theta^g\right) \right] \right\} p\left(l | x_i, \Theta^g\right)$$

Cinvestav

We can use properties of probability

We know that

$$\sum_{y_i=1}^{M} p\left(y_i|x_i,\Theta^g\right) = 1 \tag{37}$$

We can use properties of probability

We know that

$$\sum_{y_i=1}^{M} p\left(y_i|x_i, \Theta^g\right) = 1 \tag{37}$$

Then

$$\left\{ \prod_{j=1, j\neq i}^{N} \left[\sum_{y_j=1}^{M} p\left(y_j|x_j, \Theta^g\right) \right] \right\} p\left(l|x_i, \Theta^g\right) =$$

$$= \left\{ \prod_{j=1, j\neq i}^{N} 1 \right\} p\left(l|x_i, \Theta^g\right)$$

$$= p\left(l|x_i, \Theta^g\right)$$

$$= \frac{\alpha_l^g p_{y_i}\left(x_i|\theta_l^g\right)}{\sum_{k=1}^{M} \alpha_k^g p_k\left(x_i|\theta_k^g\right)}$$

We can use properties of probability

We know that

$$\sum_{y_i=1}^{M} p\left(y_i|x_i, \Theta^g\right) = 1 \tag{37}$$

Then

$$\begin{aligned}
&\left\{ \prod_{j=1, j\neq i}^{N} \left[\sum_{y_j=1}^{M} p\left(y_j|x_j, \Theta^g\right) \right] \right\} p\left(l|x_i, \Theta^g\right) = \\
&= \left\{ \prod_{j=1, j\neq i}^{N} 1 \right\} p\left(l|x_i, \Theta^g\right) \\
&= p\left(l|x_i, \Theta^g\right) \\
&= \frac{\alpha_l^g p_{y_i}\left(x_i|\theta_l^g\right)}{\sum_{k=1}^{M} \alpha_k^g p_k\left(x_i|\theta_k^g\right)}
\end{aligned}$$

We can use properties of probability

We know that

$$\sum_{y_i=1}^{M} p\left(y_i|x_i,\Theta^g\right) = 1 \tag{37}$$

Then

$$\begin{aligned}
&\left\{\prod_{j=1,j\neq i}^{N}\left[\sum_{y_j=1}^{M} p\left(y_j|x_j,\Theta^g\right)\right]\right\} p\left(l|x_i,\Theta^g\right) = \\
=&\left\{\prod_{j=1,j\neq i}^{N} 1\right\} p\left(l|x_i,\Theta^g\right)\\
=&p\left(l|x_i,\Theta^g\right)\\
=&\frac{\alpha_l^g p_{y_i}\left(x_i|\theta_l^g\right)}{\sum_{k=1}^{M}\alpha_k^g p_k\left(x_i|\theta_k^g\right)}
\end{aligned}$$

Thus

We can write Q in the following way

$$Q\left(\Theta,\Theta^{g}\right)=\sum_{i=1}^{N}\sum_{l=1}^{M}\log\left[\alpha_{l}p_{l}\left(x_{i}|\theta_{l}\right)\right]p\left(l|x_{i},\Theta^{g}\right)$$

$$=\sum_{i=1}^{N}\sum_{l=1}^{M}\log\left(\alpha_{l}\right)p\left(l|x_{i},\Theta^{g}\right)+\dots$$

$$\sum_{i=1}^{N}\sum_{l=1}^{M}\log\left(p_{l}\left(x_{i}|\theta_{l}\right)\right)p\left(l|x_{i},\Theta^{g}\right) \quad (38)$$

Thus

We can write Q in the following way

$$
\begin{aligned}
Q\left(\Theta, \Theta^{g}\right) &= \sum_{i=1}^{N}\sum_{l=1}^{M}\log\left[\alpha_l p_l\left(x_i|\theta_l\right)\right] p\left(l|x_i, \Theta^{g}\right) \\
&= \sum_{i=1}^{N}\sum_{l=1}^{M}\log\left(\alpha_l\right) p\left(l|x_i, \Theta^{g}\right) + \ldots \\
&\quad \sum_{i=1}^{N}\sum_{l=1}^{M}\log\left(p_l\left(x_i|\theta_l\right)\right) p\left(l|x_i, \Theta^{g}\right) \qquad (38)
\end{aligned}
$$

Thus

We can write Q in the following way

$$
\begin{aligned}
Q\left(\Theta,\Theta^{g}\right) &= \sum_{i=1}^{N}\sum_{l=1}^{M}\log\left[\alpha_{l}p_{l}\left(x_{i}|\theta_{l}\right)\right]p\left(l|x_{i},\Theta^{g}\right)\\
&= \sum_{i=1}^{N}\sum_{l=1}^{M}\log\left(\alpha_{l}\right)p\left(l|x_{i},\Theta^{g}\right)+...\\
&\sum_{i=1}^{N}\sum_{l=1}^{M}\log\left(p_{l}\left(x_{i}|\theta_{l}\right)\right)p\left(l|x_{i},\Theta^{g}\right)
\end{aligned} \tag{38}
$$

Outline

A Method

That could be used as a general framework

To solve problems set as EM problem.

First, we will look at the Lagrange Multipliers setup

Then, we will look at a specific case using the mixture of Gaussian's

Note

Not all the mixture of distributions will get you an analytical solution.

A Method

That could be used as a general framework

To solve problems set as EM problem.

First, we will look at the Lagrange Multipliers setup

Then, we will look at a specific case using the mixture of Gaussian's

Note

Not all the mixture of distributions will get you an analytical solution.

Cinvestav

A Method

That could be used as a general framework

To solve problems set as EM problem.

First, we will look at the Lagrange Multipliers setup

Then, we will look at a specific case using the mixture of Gaussian's

Note

Not all the mixture of distributions will get you an analytical solution.

Outline

Cinvestav

Lagrange Multipliers

The method of Lagrange multipliers

- It gives a set of necessary conditions to identify optimal points of equality constrained optimization problems.

This is done by converting a constrained problem to an equivalent unconstrained problem

- with the help of certain unspecified parameters known as Lagrange multipliers.

Lagrange Multipliers

The method of Lagrange multipliers

- It gives a set of necessary conditions to identify optimal points of equality constrained optimization problems.

This is done by converting a constrained problem to an equivalent unconstrained problem

- with the help of certain unspecified parameters known as Lagrange multipliers.

Cinvestav

Lagrange Multipliers

The classical problem formulation

$$\min \ f\left(x_1, ..., x_n\right)$$
$$s.t \ h_1\left(x_1, ..., x_n\right) = 0$$

It can be converted into

$$\min \ L\left(x_1, ..., x_n, \lambda\right) = \min\left\{f\left(x_1, ..., x_n\right) - \lambda h_1\left(x_1, ..., x_n\right)\right\}$$

where

- $L(\mathrm{x}, \lambda)$ is the Lagrangian function.
- λ is an unspecified positive or negative constant called the **Lagrange Multiplier**.

Lagrange Multipliers

The classical problem formulation

$$\min\ f\left(x_1,...,x_n\right)$$
$$s.t\ h_1\left(x_1,...,x_n\right)=0$$

It can be converted into

$$\min\ \ L\left(x_1,...,x_n,\lambda\right)=\min\left\{f\left(x_1,...,x_n\right)-\lambda h_1\left(x_1,...,x_n\right)\right\}$$

where

- $L(x,\lambda)$ is the Lagrangian function.
- λ is an unspecified positive or negative constant called the **Lagrange Multiplier**.

Lagrange Multipliers

The classical problem formulation

$$\min \; f(x_1, ..., x_n)$$
$$s.t \; h_1(x_1, ..., x_n) = 0$$

It can be converted into

$$\min \; L(x_1, ..., x_n, \lambda) = \min \{ f(x_1, ..., x_n) - \lambda h_1(x_1, ..., x_n) \}$$

where

- $L(\mathbf{x}, \lambda)$ is the Lagrangian function.
- λ is an unspecified positive or negative constant called the **Lagrange Multiplier.**

Cinvestav

Finding an Optimum using Lagrange Multipliers

New problem

$$\min \ L(x_1, ..., x_n, \lambda) = \min \{f(x_1, ..., x_n) - \lambda h_1(x_1, ..., x_n)\}$$

We want a $\lambda = \lambda^*$ optimal

If the minimum of $L(x_1, ..., x_n, \lambda^*)$ occurs at

$$(x_1, x_2, ..., x_n)^T = (x_1, x_2, ..., x_n)^{T*}$$

Cinvestav

Finding an Optimum using Lagrange Multipliers

New problem

$$\min \ L(x_1, ..., x_n, \lambda) = \min \{f(x_1, ..., x_n) - \lambda h_1(x_1, ..., x_n)\}$$

We want a $\lambda = \lambda^*$ optimal

If the minimum of $L(x_1, ..., x_n, \lambda^*)$ occurs at

$$(x_1, x_2, ..., x_n)^T = (x_1, x_2, ..., x_n)^{T*}$$

Cinvestav

Therefore

$(x_1, \ldots, x_n)^{T*}$ satisfies $h_1(x_1, \ldots, x_n) = 0$, then $(x_1, \ldots, x_n)^{T*}$ minimizes

$$\begin{aligned} &\min \ f(x_1, \ldots, x_n) \\ &s.t \ h_1(x_1, \ldots, x_n) = 0 \end{aligned}$$

Trick

- It is to find appropriate value for Lagrangian multiplier λ.

Therefore

$(x_1, ..., x_n)^{T*}$ satisfies $h_1(x_1, ..., x_n) = 0$, then $(x_1, ..., x_n)^{T*}$ minimizes

$$\begin{aligned} \min \ & f(x_1, ..., x_n) \\ s.t \ & h_1(x_1, ..., x_n) = 0 \end{aligned}$$

Trick

- It is to find appropriate value for Lagrangian multiplier λ.

Remember

Think about this

Remember First Law of Newton!!!

Yes!!!

100

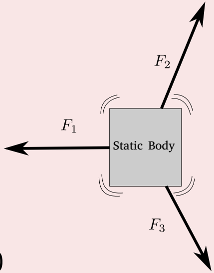
Remember

Think about this

Remember First Law of Newton!!!

Yes!!!

A system in equilibrium does not move



100

Lagrange Multipliers

Definition

Gives a set of necessary conditions to identify optimal points of equality constrained optimization problem

Lagrange was a Physicists

He was thinking in the following formula

A system in equilibrium has the following equation:

$$F_1 + F_2 + ... + F_K = 0 \tag{39}$$

But functions do not have forces?

Are you sure?

Think about the following

The Gradient of a surface.

Lagrange was a Physicists

He was thinking in the following formula

A system in equilibrium has the following equation:

$$F_1 + F_2 + ... + F_K = 0 \tag{39}$$

But functions do not have forces?

Are you sure?

Think about the following

The Gradient of a surface.

Lagrange was a Physicists

He was thinking in the following formula

A system in equilibrium has the following equation:

$$F_1 + F_2 + ... + F_K = 0 \tag{39}$$

But functions do not have forces?

Are you sure?

Think about the following

The Gradient of a surface.

Gradient to a Surface

After all a gradient is a measure of the maximal change

For example the gradient of a function of three variables:

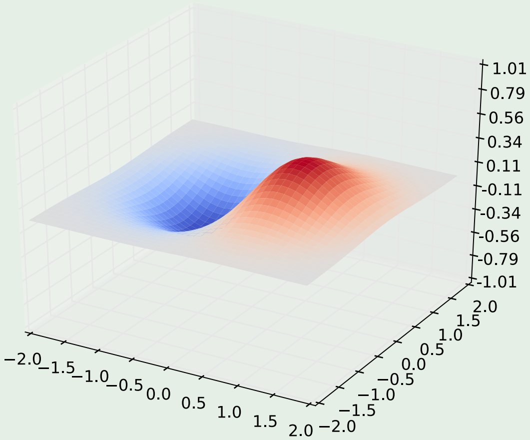
$$\nabla f(\boldsymbol{x}) = i\frac{\partial f(\boldsymbol{x})}{\partial x} + j\frac{\partial f(\boldsymbol{x})}{\partial y} + k\frac{\partial f(\boldsymbol{x})}{\partial z} \tag{40}$$

where i, j and k are unitary vectors in the directions x, y and z.

Cinvestav

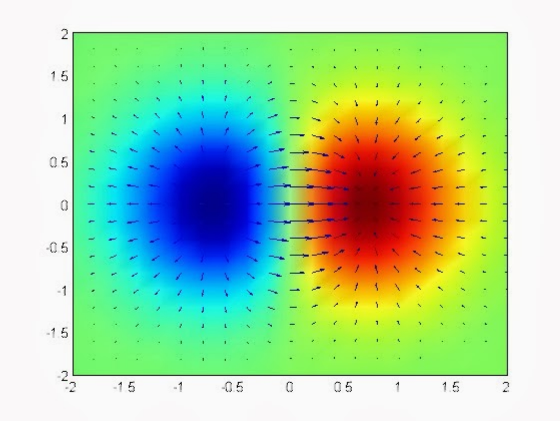
Example

We have $f(x,y) = x\exp\left\{-x^2 - y^2\right\}$



Example

With Gradient at the the contours when projecting in the 2D plane



100

Now, Think about this

Yes, we can use the gradient

However, we need to do some scaling of the forces by using parameters λ

Thus, we have

$$F_0 + \lambda_1 F_1 + ... + \lambda_K F_K = 0 \tag{41}$$

where F_0 is the gradient of the principal cost function and F_i for $i = 1, 2, ..., K$.

Now, Think about this

Yes, we can use the gradient

However, we need to do some scaling of the forces by using parameters λ

Thus, we have

$$F_0 + \lambda_1 F_1 + ... + \lambda_K F_K = 0 \tag{41}$$

where F_0 is the gradient of the principal cost function and F_i for $i = 1, 2, .., K$.

Thus

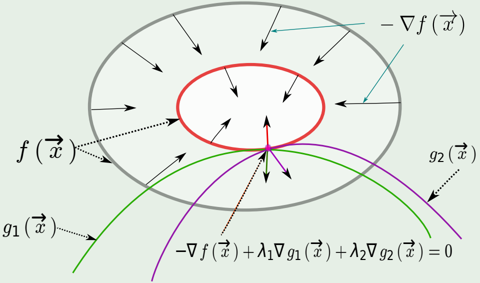
If we have the following optimization:

$$\begin{aligned} \min\ & f(\boldsymbol{x}) \\ s.t g_1(\boldsymbol{x}) &= 0 \\ g_2(\boldsymbol{x}) &= 0 \end{aligned}$$

Geometric interpretation in the case of minimization

What is wrong? Gradients are going in the other direction, we can fix by simple multiplying by -1

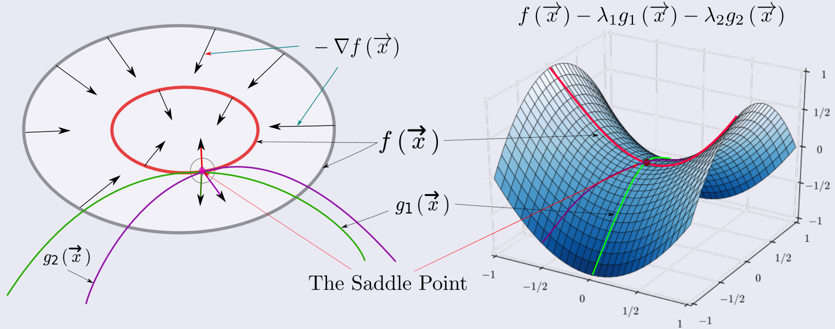
Here the cost function is $f(x,y) = x\exp\{-x^2 - y^2\}$ we want to minimize



Nevertheless: it is equivalent to $\nabla f(\overrightarrow{x}) - \lambda_1 \nabla g_1(\overrightarrow{x}) - \lambda_2 \nabla g_2(\overrightarrow{x}) = 0$

Basically, we convert the problem into a one looking for a **Saddle Point**

At the left the original problem, at the right the Lagrangian!!!



Yes!!!

Basically

- We convert the minimization or maximization of a convex or concave section of a function living in a constrained environment!!!

Method

The Basic Method

Steps

1. Original problem is rewritten as:
   1. minimize $L(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) - \lambda h_1(\boldsymbol{x})$
2. Take derivatives of $L(\boldsymbol{x}, \lambda)$ with respect to x_i and set them equal to zero.

$$\sum_{i=1}^{N} \left(y_i - \boldsymbol{x}^T \boldsymbol{w} \right)^2 + \lambda \sum_{i=1}^{d} |w_i| \tag{42}$$

3. Express all x_i in terms of Lagrangian multiplier λ.
4. Plug $\boldsymbol{x}$ in terms of λ in constraint $h_1(\boldsymbol{x}) = 0$ and solve λ.
5. Calculate $\boldsymbol{x}$ by using the just found value for λ.

From the step 2

If there are n variables (i.e., $x_1, \cdots, x_n$) then you will get n equations with $n+1$ unknowns (i.e., n variables x_i and one Lagrangian multiplier λ).

Steps

1. Original problem is rewritten as:
    1. minimize $L(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) - \lambda h_1(\boldsymbol{x})$
2. Take derivatives of $L(\boldsymbol{x}, \lambda)$ with respect to x_i and set them equal to zero.

$$\sum_{i=1}^{N} \left(y_i - \boldsymbol{x}^T \boldsymbol{w} \right)^2 + \lambda \sum_{i=1}^{d} |w_i| \tag{42}$$

3. Express all x_i in terms of Lagrangian multiplier λ.
4. Plug $\boldsymbol{x}$ in terms of λ in constraint $h_1(\boldsymbol{x}) = 0$ and solve λ.
5. Calculate $\boldsymbol{x}$ by using the just found value for λ.

From the step 2

If there are n variables (i.e., $x_1, \cdots, x_n$) then you will get n equations with $n + 1$ unknowns (i.e., n variables x_i and one Lagrangian multiplier λ).

Method

The Basic Method

Steps

1. Original problem is rewritten as:
   1. minimize $L(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) - \lambda h_1(\boldsymbol{x})$
2. Take derivatives of $L(\boldsymbol{x}, \lambda)$ with respect to x_i and set them equal to zero.
$$\sum_{i=1}^{N} \left(y_i - \boldsymbol{x}^T \boldsymbol{w} \right)^2 + \lambda \sum_{i=1}^{d} |w_i| \tag{42}$$
3. Express all x_i in terms of Lagrangian multiplier λ.
4. Plug $\boldsymbol{x}$ in terms of λ in constraint $h_1(\boldsymbol{x}) = 0$ and solve λ.
5. Calculate $\boldsymbol{x}$ by using the just found value for λ.

From the step 2

If there are n variables (i.e., $x_1, \cdots, x_n$) then you will get n equations with $n+1$ unknowns (i.e., n variables x_i and one Lagrangian multiplier λ).

Method

The Basic Method

Steps

1. Original problem is rewritten as:
   1. minimize $L(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) - \lambda h_1(\boldsymbol{x})$
2. Take derivatives of $L(\boldsymbol{x}, \lambda)$ with respect to x_i and set them equal to zero.
$$\sum_{i=1}^{N} \left(y_i - \boldsymbol{x}^T \boldsymbol{w} \right)^2 + \lambda \sum_{i=1}^{d} |w_i| \tag{42}$$
3. Express all x_i in terms of Lagrangian multiplier λ.
4. Plug $\boldsymbol{x}$ in terms of λ in constraint $h_1(\boldsymbol{x}) = 0$ and solve λ.
5. Calculate x by using the just found value for λ.

From the step 2

If there are n variables (i.e., $x_1, \cdots, x_n$) then you will get n equations with $n+1$ unknowns (i.e., n variables x_i and one Lagrangian multiplier λ).

Method

The Basic Method

Steps

1. Original problem is rewritten as:
   1. minimize $L(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) - \lambda h_1(\boldsymbol{x})$
2. Take derivatives of $L(\boldsymbol{x}, \lambda)$ with respect to x_i and set them equal to zero.

$$\sum_{i=1}^{N} \left(y_i - \boldsymbol{x}^T \boldsymbol{w}\right)^2 + \lambda \sum_{i=1}^{d} |w_i| \tag{42}$$

3. Express all x_i in terms of Lagrangian multiplier λ.
4. Plug $\boldsymbol{x}$ in terms of λ in constraint $h_1(\boldsymbol{x}) = 0$ and solve λ.
5. Calculate $\boldsymbol{x}$ by using the just found value for λ.

From the step 2

If there are n variables (i.e., $x_1, \cdots, x_n$) then you will get n equations with $n+1$ unknowns (i.e., n variables x_i and one Lagrangian multiplier λ).

Method

The Basic Method

Steps

1. Original problem is rewritten as:
   1. minimize $L(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) - \lambda h_1(\boldsymbol{x})$
2. Take derivatives of $L(\boldsymbol{x}, \lambda)$ with respect to x_i and set them equal to zero.

$$\sum_{i=1}^{N}\left(y_i - \boldsymbol{x}^T\boldsymbol{w}\right)^2 + \lambda\sum_{i=1}^{d}|w_i| \tag{42}$$

3. Express all x_i in terms of Lagrangian multiplier λ.
4. Plug $\boldsymbol{x}$ in terms of λ in constraint $h_1(\boldsymbol{x}) = 0$ and solve λ.
5. Calculate $\boldsymbol{x}$ by using the just found value for λ.

From the step 2

If there are n variables (i.e., $x_1, \cdots, x_n$) then you will get n equations with $n+1$ unknowns (i.e., n variables x_i and one Lagrangian multiplier λ).

Example

We can apply that to the following problem

$$
\begin{aligned}
\min \ & f(x, y) = x^2 - 8x + y^2 - 12y + 48 \\
s.t \ & x + y = 8
\end{aligned}
$$

Outline

Cinvestav

Lagrange Multipliers for Q

We can us the following constraint for that

$$\sum_l \alpha_l = 1 \tag{43}$$

We have the following cost function

$$Q\left(\Theta, \Theta^g\right) + \lambda\left(\sum_l \alpha_l - 1\right) \tag{44}$$

Deriving by α_l

$$\frac{\partial}{\partial \alpha_l}\left[Q\left(\Theta, \Theta^g\right) + \lambda\left(\sum_l \alpha_l - 1\right)\right] = 0 \tag{45}$$

Lagrange Multipliers for Q

We can us the following constraint for that

$$\sum_l \alpha_l = 1 \tag{43}$$

We have the following cost function

$$Q\left(\Theta, \Theta^g\right) + \lambda \left(\sum_l \alpha_l - 1\right) \tag{44}$$

Deriving by α_l

$$\frac{\partial}{\partial \alpha_l}\left[Q\left(\Theta, \Theta^g\right) + \lambda \left(\sum_l \alpha_l - 1\right)\right] = 0 \tag{45}$$

Lagrange Multipliers for Q

We can us the following constraint for that

$$\sum_l \alpha_l = 1 \tag{43}$$

We have the following cost function

$$Q\left(\Theta, \Theta^g\right) + \lambda \left(\sum_l \alpha_l - 1 \right) \tag{44}$$

Deriving by α_l

$$\frac{\partial}{\partial \alpha_l} \left[Q\left(\Theta, \Theta^g\right) + \lambda \left(\sum_l \alpha_l - 1 \right) \right] = 0 \tag{45}$$

Thus

The Q function

$$Q\left(\Theta,\Theta^{g}\right)=\sum_{i=1}^{N}\sum_{l=1}^{M}\log\left(\alpha_{l}\right)p\left(l|x_{i},\Theta^{g}\right)+...$$

$$\sum_{i=1}^{N}\sum_{l=1}^{M}\log\left(p_{l}\left(x_{i}|\theta_{l}\right)\right)p\left(l|x_{i},\Theta^{g}\right)$$

Cinvestav

Deriving

We have

$$\frac{\partial}{\partial \alpha_l}\left[Q\left(\Theta,\Theta^g\right)+\lambda\left(\sum_l \alpha_l-1\right)\right]=\sum_{i=1}^{N}\frac{1}{\alpha_l}p\left(l|x_i,\Theta^g\right)+\lambda$$

Finally

We have making the previous equation equal to 0

$$\sum_{i=1}^{N} \frac{1}{\alpha_l} p\left(l|x_i, \Theta^g\right) + \lambda = 0 \quad (46)$$

Thus

$$\sum_{i=1}^{N} p\left(l|x_i, \Theta^g\right) = -\lambda \alpha_l \quad (47)$$

Summing over l, we get

$$\lambda = -N \quad (48)$$

Finally

We have making the previous equation equal to 0

$$\sum_{i=1}^{N} \frac{1}{\alpha_l} p\left(l|x_i, \Theta^g\right) + \lambda = 0 \tag{46}$$

Thus

$$\sum_{i=1}^{N} p\left(l|x_i, \Theta^g\right) = -\lambda\alpha_l \tag{47}$$

Summing over l, we get

$$\lambda = -N \tag{48}$$

Cinvestav

Finally

We have making the previous equation equal to 0

$$\sum_{i=1}^{N} \frac{1}{\alpha_l} p\left(l|x_i, \Theta^g\right) + \lambda = 0 \tag{46}$$

Thus

$$\sum_{i=1}^{N} p\left(l|x_i, \Theta^g\right) = -\lambda \alpha_l \tag{47}$$

Summing over l, we get

$$\lambda = -N \tag{48}$$

Lagrange Multipliers

Thus

$$\alpha_l = \frac{1}{N} \sum_{i=1}^{N} p\left(l|x_i, \Theta^g\right) \tag{49}$$

About θ_l

It is possible to get an analytical expressions for θ_l as functions of everything else.

- This is for you to try!!!

For more, please look at

"Geometric Idea of Lagrange Multipliers" by John Wyatt.

Cinvestav

Lagrange Multipliers

Thus

$$\alpha_l = \frac{1}{N} \sum_{i=1}^{N} p\left(l|x_i, \Theta^g\right) \quad (49)$$

About θ_l

It is possible to get an analytical expressions for θ_l as functions of everything else.

- This is for you to try!!!

For more, please look at

"Geometric Idea of Lagrange Multipliers" by John Wyatt.

Cinvestav

Lagrange Multipliers

Thus

$$\alpha_l = \frac{1}{N} \sum_{i=1}^{N} p\left(l|x_i, \Theta^g\right) \tag{49}$$

About θ_l

It is possible to get an analytical expressions for θ_l as functions of everything else.

- This is for you to try!!!

For more, please look at

"Geometric Idea of Lagrange Multipliers" by John Wyatt.

Cinvestav

Outline

Cinvestav

Remember?

Gaussian Distribution

$$p_l\left(\boldsymbol{x}|\boldsymbol{\mu}_l,\boldsymbol{\Sigma}_l\right)=\frac{1}{\left(2\pi\right)^{d/2}\left|\Sigma_l\right|^{1/2}}\exp\left\{-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}_l\right)^T\Sigma_l^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}_l\right)\right\} \quad (50)$$

Cinvestav

How to use this for Gaussian Distributions

For this, we need to refresh some linear algebra

1. $tr\left(A+B\right)=tr\left(A\right)+tr\left(B\right)$
2. $tr\left(AB\right)=tr\left(BA\right)$
3. $\sum_i x_i^T A x_i = tr\left(AB\right)$ where $B=\sum_i x_i x_i^T$.
4. $\left|A^{-1}\right| = \frac{1}{|A|}$

Now, we need the derivative of a matrix function $f\left(A\right)$

Thus, $\frac{\partial f(A)}{\partial A}$ is going to be the matrix with i,j^{th} entry $\left[\frac{\partial f(A)}{\partial a_{i,j}}\right]$ where $a_{i,j}$ is the i,j^{th} entry of A.

How to use this for Gaussian Distributions

For this, we need to refresh some linear algebra

1. $tr\left(A+B\right)=tr\left(A\right)+tr\left(B\right)$
2. $tr\left(AB\right)=tr\left(BA\right)$
3. $\sum_i x_i^T A x_i = tr\left(AB\right)$ where $B=\sum_i x_i x_i^T$.
4. $\left|A^{-1}\right| = \frac{1}{|A|}$

Now, we need the derivative of a matrix function $f\left(A\right)$

Thus, $\frac{\partial f(A)}{\partial A}$ is going to be the matrix with i,j^{th} entry $\left[\frac{\partial f(A)}{\partial a_{i,j}}\right]$ where $a_{i,j}$ is the i,j^{th} entry of A.

How to use this for Gaussian Distributions

For this, we need to refresh some linear algebra

1. $tr\left(A+B\right)=tr\left(A\right)+tr\left(B\right)$
2. $tr\left(AB\right)=tr\left(BA\right)$
3. $\sum_i x_i^T A x_i = tr\left(AB\right)$ where $B=\sum_i x_i x_i^T$.
4. $\left|A^{-1}\right| = \frac{1}{|A|}$

Now, we need the derivative of a matrix function $f\left(A\right)$

Thus, $\frac{\partial f(A)}{\partial A}$ is going to be the matrix with i,j^{th} entry $\left[\frac{\partial f(A)}{\partial a_{i,j}}\right]$ where $a_{i,j}$ is the i,j^{th} entry of A.

How to use this for Gaussian Distributions

For this, we need to refresh some linear algebra

1. $tr(A+B) = tr(A) + tr(B)$
2. $tr(AB) = tr(BA)$
3. $\sum_i x_i^T A x_i = tr(AB)$ where $B = \sum_i x_i x_i^T$.
4. $\left|A^{-1}\right| = \frac{1}{|A|}$

Now, we need the derivative of a matrix function $f(A)$

Thus, $\frac{\partial f(A)}{\partial A}$ is going to be the matrix with i, j^{th} entry $\left[\frac{\partial f(A)}{\partial a_{i,j}}\right]$ where $a_{i,j}$ is the i, j^{th} entry of A.

Cinvestav

How to use this for Gaussian Distributions

For this, we need to refresh some linear algebra

1. $tr\left(A+B\right)=tr\left(A\right)+tr\left(B\right)$
2. $tr\left(AB\right)=tr\left(BA\right)$
3. $\sum_i x_i^T A x_i = tr\left(AB\right)$ where $B=\sum_i x_i x_i^T$.
4. $\left|A^{-1}\right| = \frac{1}{|A|}$

Now, we need the derivative of a matrix function $f\left(A\right)$

Thus, $\frac{\partial f(A)}{\partial A}$ is going to be the matrix with i,j^{th} entry $\left[\frac{\partial f(A)}{\partial a_{i,j}}\right]$ where $a_{i,j}$ is the i,j^{th} entry of A.

In addition

If A is symmetric

$$\frac{\partial |A|}{\partial A} = \begin{cases} \mathcal{A}_{i,j} & \text{if } i = j \\ 2\mathcal{A}_{i,j} & \text{if } i \neq j \end{cases} \tag{51}$$

Where $\mathcal{A}_{i,j}$ is the i, j^{th} cofactor of A.

Note: The determinant obtained by deleting the row and column of a given element of a matrix or determinant. The **cofactor** is preceded by a $+$ or $-$ sign depending whether the element is in a $+$ or $-$ position.

Thus

$$\frac{\partial \log |A|}{\partial A} = \begin{cases} \frac{\mathcal{A}_{i,j}}{|A|} & \text{if } i = j \\ 2\mathcal{A}_{i,j} & \text{if } i \neq j \end{cases} = 2A^{-1} - \text{diag}\left(A^{-1}\right) \tag{52}$$

In addition

If A is symmetric

$$\frac{\partial |A|}{\partial A} = \begin{cases} \mathcal{A}_{i,j} & \text{if } i = j \\ 2\mathcal{A}_{i,j} & \text{if } i \neq j \end{cases} \tag{51}$$

Where $\mathcal{A}_{i,j}$ is the i, j^{th} cofactor of A.

Note: The determinant obtained by deleting the row and column of a given element of a matrix or determinant. The **cofactor** is preceded by a $+$ or $-$ sign depending whether the element is in a $+$ or $-$ position.

Thus

$$\frac{\partial \log |A|}{\partial A} = \begin{cases} \frac{\mathcal{A}_{i,j}}{|A|} & \text{if } i = j \\ 2\mathcal{A}_{i,j} & \text{if } i \neq j \end{cases} = 2A^{-1} - \text{diag}\left(A^{-1}\right) \tag{52}$$

In addition

If A is symmetric

$$\frac{\partial |A|}{\partial A} = \begin{cases} \mathcal{A}_{i,j} & \text{if } i = j \\ 2\mathcal{A}_{i,j} & \text{if } i \neq j \end{cases} \tag{51}$$

Where $\mathcal{A}_{i,j}$ is the i, j^{th} cofactor of A.

Note: The determinant obtained by deleting the row and column of a given element of a matrix or determinant. The **cofactor** is preceded by a $+$ or – sign depending whether the element is in a $+$ or – position.

Thus

$$\frac{\partial \log |A|}{\partial A} = \begin{cases} \frac{\mathcal{A}_{i,j}}{|A|} & \text{if } i = j \\ 2\mathcal{A}_{i,j} & \text{if } i \neq j \end{cases} = 2A^{-1} - \text{diag}\left(A^{-1}\right) \tag{52}$$

In addition

If A is symmetric

$$\frac{\partial |A|}{\partial A} = \begin{cases} \mathcal{A}_{i,j} & \text{if } i = j \\ 2\mathcal{A}_{i,j} & \text{if } i \neq j \end{cases} \tag{51}$$

Where $\mathcal{A}_{i,j}$ is the i, j^{th} cofactor of A.

Note: The determinant obtained by deleting the row and column of a given element of a matrix or determinant. The **cofactor** is preceded by a $+$ or – sign depending whether the element is in a $+$ or – position.

Thus

$$\frac{\partial \log |A|}{\partial A} = \begin{cases} \frac{\mathcal{A}_{i,j}}{|A|} & \text{if } i = j \\ 2\mathcal{A}_{i,j} & \text{if } i \neq j \end{cases} = 2A^{-1} - \text{diag}\left(A^{-1}\right) \tag{52}$$

Finally

The last equation we need

$$\frac{\partial tr\left(AB\right)}{\partial A} = B + B^T - \mathrm{diag}\left(B\right) \tag{53}$$

In addition

$$\frac{\partial x^T A x}{\partial x} \tag{54}$$

Finally

The last equation we need

$$\frac{\partial tr\left(AB\right)}{\partial A} = B + B^T - \mathrm{diag}\left(B\right) \quad (53)$$

In addition

$$\frac{\partial \boldsymbol{x}^T A \boldsymbol{x}}{\partial \boldsymbol{x}} \quad (54)$$

Thus, using last part of equation 38

We get, after ignoring constant terms

Remember they disappear after derivatives

$$\sum_{i=1}^{N}\sum_{l=1}^{M}\log\left(p_l\left(\boldsymbol{x}_i|\mu_l,\Sigma_l\right)\right)p\left(l|\boldsymbol{x}_i,\Theta^g\right)$$

$$=\sum_{i=1}^{N}\sum_{l=1}^{M}\left[-\frac{1}{2}\log\left(|\Sigma_l|\right)-\frac{1}{2}\left(\boldsymbol{x}_i-\mu_l\right)^T\Sigma_l^{-1}\left(\boldsymbol{x}_i-\mu_l\right)\right]p\left(l|\boldsymbol{x}_i,\Theta^g\right) \quad (55)$$

Cinvestav

Thus, using last part of equation 38

We get, after ignoring constant terms

Remember they disappear after derivatives

$$
\begin{aligned}
&\sum_{i=1}^{N} \sum_{l=1}^{M} \log\left(p_l\left(\boldsymbol{x}_i|\mu_l, \Sigma_l\right)\right) p\left(l|\boldsymbol{x}_i, \Theta^g\right) \\
&= \sum_{i=1}^{N} \sum_{l=1}^{M} \left[-\frac{1}{2} \log\left(|\Sigma_l|\right) - \frac{1}{2}\left(\boldsymbol{x}_i - \mu_l\right)^T \Sigma_l^{-1}\left(\boldsymbol{x}_i - \mu_l\right)\right] p\left(l|\boldsymbol{x}_i, \Theta^g\right) \quad (55)
\end{aligned}
$$

Finally

Thus, when taking the derivative with respect to μ_l

$$\sum_{i=1}^{N}\left[\Sigma_l^{-1}\left(\boldsymbol{x}_i-\mu_l\right)p\left(l|\boldsymbol{x}_i,\Theta^g\right)\right]=0 \tag{56}$$

Then

$$\mu_l=\frac{\sum_{i=1}^{N}\boldsymbol{x}_i p\left(l|\boldsymbol{x}_i,\Theta^g\right)}{\sum_{i=1}^{N}p\left(l|\boldsymbol{x}_i,\Theta^g\right)} \tag{57}$$

Finally

Thus, when taking the derivative with respect to μ_l

$$\sum_{i=1}^{N} \left[\Sigma_l^{-1} \left(\boldsymbol{x}_i - \mu_l \right) p \left(l | \boldsymbol{x}_i, \Theta^g \right) \right] = 0 \tag{56}$$

Then

$$\mu_l = \frac{\sum_{i=1}^{N} \boldsymbol{x}_i p \left(l | \boldsymbol{x}_i, \Theta^g \right)}{\sum_{i=1}^{N} p \left(l | \boldsymbol{x}_i, \Theta^g \right)} \tag{57}$$

Now, if we derive with respect to Σ_l

First, we rewrite equation 55

$$\sum_{i=1}^{N}\sum_{l=1}^{M}\left[-\frac{1}{2}\log\left(|\Sigma_l|\right)-\frac{1}{2}\left(\boldsymbol{x}_i-\mu_l\right)^T\Sigma_l^{-1}\left(\boldsymbol{x}_i-\mu_l\right)\right]p\left(l|\boldsymbol{x}_i,\Theta^g\right)$$

$$=\sum_{l=1}^{M}\left[-\frac{1}{2}\log\left(|\Sigma_l|\right)\sum_{i=1}^{N}p\left(l|\boldsymbol{x}_i,\Theta^g\right)-\frac{1}{2}\sum_{i=1}^{N}p\left(l|\boldsymbol{x}_i,\Theta^g\right)tr\left\{\Sigma_l^{-1}\left(\boldsymbol{x}_i-\mu_l\right)\left(\boldsymbol{x}_i-\mu_l\right)^T\right\}\right]$$

$$=\sum_{l=1}^{M}\left[-\frac{1}{2}\log\left(|\Sigma_l|\right)\sum_{i=1}^{N}p\left(l|\boldsymbol{x}_i,\Theta^g\right)-\frac{1}{2}\sum_{i=1}^{N}p\left(l|\boldsymbol{x}_i,\Theta^g\right)tr\left\{\Sigma_l^{-1}N_{l,i}\right\}\right]$$

Where $N_{l,i}=\left(\boldsymbol{x}_i-\mu_l\right)\left(\boldsymbol{x}_i-\mu_l\right)^T$.

Now, if we derive with respect to Σ_l

First, we rewrite equation 55

$$\sum_{i=1}^{N}\sum_{l=1}^{M}\left[-\frac{1}{2}\log\left(|\Sigma_l|\right)-\frac{1}{2}\left(\boldsymbol{x}_i-\mu_l\right)^T\Sigma_l^{-1}\left(\boldsymbol{x}_i-\mu_l\right)\right]p\left(l|\boldsymbol{x}_i,\Theta^g\right)$$

$$=\sum_{l=1}^{M}\left[-\frac{1}{2}\log\left(|\Sigma_l|\right)\sum_{i=1}^{N}p\left(l|\boldsymbol{x}_i,\Theta^g\right)-\frac{1}{2}\sum_{i=1}^{N}p\left(l|\boldsymbol{x}_i,\Theta^g\right)tr\left\{\Sigma_l^{-1}\left(\boldsymbol{x}_i-\mu_l\right)\left(\boldsymbol{x}_i-\mu_l\right)^T\right\}\right]$$

$$=\sum_{l=1}^{M}\left[-\frac{1}{2}\log\left(|\Sigma_l|\right)\sum_{i=1}^{N}p\left(l|\boldsymbol{x}_i,\Theta^g\right)-\frac{1}{2}\sum_{i=1}^{N}p\left(l|\boldsymbol{x}_i,\Theta^g\right)tr\left\{\Sigma_l^{-1}N_{l,i}\right\}\right]$$

Where $N_{l,i}=\left(\boldsymbol{x}_i-\mu_l\right)\left(\boldsymbol{x}_i-\mu_l\right)^T$.

Now, if we derive with respect to Σ_l

First, we rewrite equation 55

$$\sum_{i=1}^{N}\sum_{l=1}^{M}\left[-\frac{1}{2}\log\left(\left|\Sigma_l\right|\right)-\frac{1}{2}\left(\boldsymbol{x}_i-\mu_l\right)^T\Sigma_l^{-1}\left(\boldsymbol{x}_i-\mu_l\right)\right]p\left(l|\boldsymbol{x}_i,\Theta^g\right)$$

$$=\sum_{l=1}^{M}\left[-\frac{1}{2}\log\left(\left|\Sigma_l\right|\right)\sum_{i=1}^{N}p\left(l|\boldsymbol{x}_i,\Theta^g\right)-\frac{1}{2}\sum_{i=1}^{N}p\left(l|\boldsymbol{x}_i,\Theta^g\right)tr\left\{\Sigma_l^{-1}\left(\boldsymbol{x}_i-\mu_l\right)\left(\boldsymbol{x}_i-\mu_l\right)^T\right\}\right]$$

$$=\sum_{l=1}^{M}\left[-\frac{1}{2}\log\left(\left|\Sigma_l\right|\right)\sum_{i=1}^{N}p\left(l|\boldsymbol{x}_i,\Theta^g\right)-\frac{1}{2}\sum_{i=1}^{N}p\left(l|\boldsymbol{x}_i,\Theta^g\right)tr\left\{\Sigma_l^{-1}N_{l,i}\right\}\right]$$

Where $N_{l,i}=\left(\boldsymbol{x}_i-\mu_l\right)\left(\boldsymbol{x}_i-\mu_l\right)^T$.

Now, if we derive with respect to Σ_l

First, we rewrite equation 55

$$\sum_{i=1}^{N}\sum_{l=1}^{M}\left[-\frac{1}{2}\log\left(|\Sigma_l|\right)-\frac{1}{2}\left(\boldsymbol{x}_i-\mu_l\right)^T\Sigma_l^{-1}\left(\boldsymbol{x}_i-\mu_l\right)\right]p\left(l|\boldsymbol{x}_i,\Theta^g\right)$$

$$=\sum_{l=1}^{M}\left[-\frac{1}{2}\log\left(|\Sigma_l|\right)\sum_{i=1}^{N}p\left(l|\boldsymbol{x}_i,\Theta^g\right)-\frac{1}{2}\sum_{i=1}^{N}p\left(l|\boldsymbol{x}_i,\Theta^g\right)tr\left\{\Sigma_l^{-1}\left(\boldsymbol{x}_i-\mu_l\right)\left(\boldsymbol{x}_i-\mu_l\right)^T\right\}\right]$$

$$=\sum_{l=1}^{M}\left[-\frac{1}{2}\log\left(|\Sigma_l|\right)\sum_{i=1}^{N}p\left(l|\boldsymbol{x}_i,\Theta^g\right)-\frac{1}{2}\sum_{i=1}^{N}p\left(l|\boldsymbol{x}_i,\Theta^g\right)tr\left\{\Sigma_l^{-1}N_{l,i}\right\}\right]$$

Where $N_{l,i}=\left(\boldsymbol{x}_i-\mu_l\right)\left(\boldsymbol{x}_i-\mu_l\right)^T$.

Deriving with respect to Σ_l^{-1}

We have that

$$\frac{\partial}{\partial \Sigma_l^{-1}} \sum_{l=1}^{M} \left[-\frac{1}{2} \log\left(|\Sigma_l|\right) \sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i, \Theta^g\right) - \frac{1}{2} \sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i, \Theta^g\right) tr\left\{\Sigma_l^{-1} N_{l,i}\right\} \right]$$

$$= \frac{1}{2} \sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i, \Theta^g\right)\left(2\Sigma_l - \text{diag}\left(\Sigma_l\right)\right) - \frac{1}{2} \sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i, \Theta^g\right)\left(2N_{l,i} - \text{diag}\left(N_{l,i}\right)\right)$$

$$= \frac{1}{2} \sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i, \Theta^g\right)\left(2M_{l,i} - \text{diag}\left(M_{l,i}\right)\right)$$

$$= 2S - \text{diag}\left(S\right)$$

Where $M_{l,i} = \Sigma_l - N_{l,i}$ and $S = \frac{1}{2} \sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i, \Theta^g\right) M_{l,i}$

Deriving with respect to Σ_l^{-1}

We have that

$$\begin{aligned}
\frac{\partial}{\partial \Sigma_l^{-1}} \sum_{l=1}^{M} \left[-\frac{1}{2}\log\left(|\Sigma_l|\right) \sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i,\Theta^g\right) - \frac{1}{2}\sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i,\Theta^g\right) tr\left\{\Sigma_l^{-1} N_{l,i}\right\}\right] \\
=& \frac{1}{2}\sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i,\Theta^g\right)\left(2\Sigma_l - \text{diag}\left(\Sigma_l\right)\right) - \frac{1}{2}\sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i,\Theta^g\right)\left(2N_{l.i} - \text{diag}\left(N_{l,i}\right)\right) \\
=& \frac{1}{2}\sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i,\Theta^g\right)\left(2M_{l,i} - \text{diag}\left(M_{l,i}\right)\right) \\
=& 2S - \text{diag}\left(S\right)
\end{aligned}$$

Where $M_{l,i} = \Sigma_l - N_{l,i}$ and $S = \frac{1}{2}\sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i,\Theta^g\right) M_{l,i}$

Deriving with respect to Σ_l^{-1}

We have that

$$\frac{\partial}{\partial \Sigma_l^{-1}} \sum_{l=1}^{M} \left[-\frac{1}{2} \log\left(|\Sigma_l|\right) \sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i, \Theta^g\right) - \frac{1}{2} \sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i, \Theta^g\right) tr\left\{ \Sigma_l^{-1} N_{l,i} \right\} \right]$$

$$= \frac{1}{2} \sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i, \Theta^g\right) \left(2\Sigma_l - \text{diag}\left(\Sigma_l\right)\right) - \frac{1}{2} \sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i, \Theta^g\right) \left(2N_{l.i} - \text{diag}\left(N_{l,i}\right)\right)$$

$$= \frac{1}{2} \sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i, \Theta^g\right) \left(2M_{l,i} - \text{diag}\left(M_{l,i}\right)\right)$$

$$= 2S - \text{diag}\left(S\right)$$

Where $M_{l,i} = \Sigma_l - N_{l,i}$ and $S = \frac{1}{2}\sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i, \Theta^g\right) M_{l,i}$

Deriving with respect to Σ_l^{-1}

We have that

$$
\begin{aligned}
&\frac{\partial}{\partial \Sigma_l^{-1}} \sum_{l=1}^{M} \left[-\frac{1}{2} \log\left(|\Sigma_l|\right) \sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i, \Theta^g\right) - \frac{1}{2} \sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i, \Theta^g\right) tr\left\{ \Sigma_l^{-1} N_{l,i} \right\} \right] \\
&= \frac{1}{2} \sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i, \Theta^g\right) \left(2\Sigma_l - \text{diag}\left(\Sigma_l\right)\right) - \frac{1}{2} \sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i, \Theta^g\right) \left(2N_{l.i} - \text{diag}\left(N_{l,i}\right)\right) \\
&= \frac{1}{2} \sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i, \Theta^g\right) \left(2M_{l,i} - \text{diag}\left(M_{l,i}\right)\right) \\
&= 2S - \text{diag}\left(S\right)
\end{aligned}
$$

Where $M_{l,i} = \Sigma_l - N_{l,i}$ and $S = \frac{1}{2} \sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i, \Theta^g\right) M_{l,i}$

Cinvestav

Thus, we have

Thus

If $2S - \text{diag}(S) = 0 \Longrightarrow S = 0$

Implying

$$\frac{1}{2}\sum_{i=1}^{N} p\left(l|x_i,\Theta^g\right)\left[\Sigma_l - N_{l,i}\right] = 0 \quad (58)$$

Or

$$\Sigma_l = \frac{\sum_{i=1}^{N} p\left(l|x_i,\Theta^g\right) N_{l,i}}{\sum_{i=1}^{N} p\left(l|x_i,\Theta^g\right)} = \frac{\sum_{i=1}^{N} p\left(l|x_i,\Theta^g\right)\left(x_i-\mu_l\right)\left(x_i-\mu_l\right)^T}{\sum_{i=1}^{N} p\left(l|x_i,\Theta^g\right)} \quad (59)$$

Thus, we have

Thus

If $2S - \text{diag}(S) = 0 \Longrightarrow S = 0$

Implying

$$\frac{1}{2}\sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i, \Theta^g\right)\left[\Sigma_l - N_{l,i}\right] = 0 \tag{58}$$

Or

$$\Sigma_l = \frac{\sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i, \Theta^g\right) N_{l,i}}{\sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i, \Theta^g\right)} = \frac{\sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i, \Theta^g\right)\left(\boldsymbol{x}_i - \mu_l\right)\left(\boldsymbol{x}_i - \mu_l\right)^T}{\sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i, \Theta^g\right)} \tag{59}$$

Thus, we have

Thus

If $2S - \text{diag}(S) = 0 \Longrightarrow S = 0$

Implying

$$\frac{1}{2}\sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i, \Theta^g\right)\left[\Sigma_l - N_{l,i}\right] = 0 \tag{58}$$

Or

$$\Sigma_l = \frac{\sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i, \Theta^g\right) N_{l,i}}{\sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i, \Theta^g\right)} = \frac{\sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i, \Theta^g\right)\left(\boldsymbol{x}_i - \mu_l\right)\left(\boldsymbol{x}_i - \mu_l\right)^T}{\sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i, \Theta^g\right)} \tag{59}$$

Cinvestav

Thus, we have the iterative updates

They are

$$\alpha_l^{New} = \frac{1}{N} \sum_{i=1}^{N} p\left(l|x_i, \Theta^g\right)$$

$$\mu_l^{New} = \frac{\sum_{i=1}^{N} x_i p\left(l|x_i, \Theta^g\right)}{\sum_{i=1}^{N} p\left(l|x_i, \Theta^g\right)}$$

$$\Sigma_l^{New} = \frac{\sum_{i=1}^{N} p\left(l|x_i, \Theta^g\right)\left(x_i - \mu_l\right)\left(x_i - \mu_l\right)^T}{\sum_{i=1}^{N} p\left(l|x_i, \Theta^g\right)}$$

Cinvestav

Thus, we have the iterative updates

They are

$$
\begin{aligned}
\alpha_l^{New} &= \frac{1}{N}\sum_{i=1}^{N} p\left(l|x_i,\Theta^g\right)\\
\mu_l^{New} &= \frac{\sum_{i=1}^{N} \boldsymbol{x}_i p\left(l|\boldsymbol{x}_i,\Theta^g\right)}{\sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i,\Theta^g\right)}\\
\Sigma_l^{New} &= \frac{\sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i,\Theta^g\right)\left(\boldsymbol{x}_i-\mu_l\right)\left(\boldsymbol{x}_i-\mu_l\right)^T}{\sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i,\Theta^g\right)}
\end{aligned}
$$

Thus, we have the iterative updates

They are

$$
\begin{aligned}
\alpha_l^{New} &= \frac{1}{N}\sum_{i=1}^{N} p\left(l|x_i,\Theta^g\right)\\
\mu_l^{New} &= \frac{\sum_{i=1}^{N} \boldsymbol{x}_i p\left(l|\boldsymbol{x}_i,\Theta^g\right)}{\sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i,\Theta^g\right)}\\
\Sigma_l^{New} &= \frac{\sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i,\Theta^g\right)\left(\boldsymbol{x}_i-\mu_l\right)\left(\boldsymbol{x}_i-\mu_l\right)^T}{\sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i,\Theta^g\right)}
\end{aligned}
$$

Outline

Cinvestav

EM Algorithm for Gaussian Mixtures

Step 1

Initialize:

- The means μ_l
- Covariances Σ_l
- Mixing coefficients α_l

Evaluate

Step 2 - **E-Step**

- Evaluate the the probabilities of component l given x_i using the current parameter values:

$$p\left(l|x_i, \Theta^g\right) = \frac{\alpha_l^g p_{y_i}\left(x_i|\theta_l^g\right)}{\sum_{k=1}^{M} \alpha_k^g p_k\left(x_i|\theta_k^g\right)}$$

Now

Step 3 - **M-Step**

- Re-estimate the parameters using the current iteration values:

$$\alpha_l^{New} = \frac{1}{N}\sum_{i=1}^{N} p\left(l|x_i, \Theta^g\right)$$

$$\mu_l^{New} = \frac{\sum_{i=1}^{N} x_i p\left(l|x_i, \Theta^g\right)}{\sum_{i=1}^{N} p\left(l|x_i, \Theta^g\right)}$$

$$\Sigma_l^{New} = \frac{\sum_{i=1}^{N} p\left(l|x_i, \Theta^g\right)\left(x_i - \mu_l\right)\left(x_i - \mu_l\right)^T}{\sum_{i=1}^{N} p\left(l|x_i, \Theta^g\right)}$$

Now

Step 3 - **M-Step**

- Re-estimate the parameters using the current iteration values:

$$
\begin{aligned}
\alpha_l^{New} &= \frac{1}{N}\sum_{i=1}^{N} p\left(l|x_i,\Theta^g\right)\\
\mu_l^{New} &= \frac{\sum_{i=1}^{N} \boldsymbol{x}_i p\left(l|\boldsymbol{x}_i,\Theta^g\right)}{\sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i,\Theta^g\right)}\\
\Sigma_l^{New} &= \frac{\sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i,\Theta^g\right)\left(\boldsymbol{x}_i-\mu_l\right)\left(\boldsymbol{x}_i-\mu_l\right)^T}{\sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i,\Theta^g\right)}
\end{aligned}
$$

Now

Step 3 - **M-Step**

- Re-estimate the parameters using the current iteration values:

$$\alpha_l^{New} = \frac{1}{N}\sum_{i=1}^{N} p\left(l|x_i, \Theta^g\right)$$

$$\mu_l^{New} = \frac{\sum_{i=1}^{N} \boldsymbol{x}_i p\left(l|\boldsymbol{x}_i, \Theta^g\right)}{\sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i, \Theta^g\right)}$$

$$\Sigma_l^{New} = \frac{\sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i, \Theta^g\right)\left(\boldsymbol{x}_i - \mu_l\right)\left(\boldsymbol{x}_i - \mu_l\right)^T}{\sum_{i=1}^{N} p\left(l|\boldsymbol{x}_i, \Theta^g\right)}$$

Cinvestav

Evaluate

Step 4

Evaluate the log likelihood:

$$\log p\left(\boldsymbol{X}|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\alpha}\right)=\sum_{i=1}^{N}\log\left\{\sum_{l=1}^{M}\alpha_{l}^{New}p_{l}\left(\boldsymbol{x_i}|\boldsymbol{\mu_l^{New}},\boldsymbol{\Sigma_l}^{New}\right)\right\}$$

Step 6

- Check for convergence of either the parameters or the log likelihood.
- If the convergence criterion is not satisfied return to step 2.

Evaluate

Step 4

Evaluate the log likelihood:

$$\log p\left(\boldsymbol{X}|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\alpha}\right)=\sum_{i=1}^{N}\log\left\{\sum_{l=1}^{M}\alpha_{l}^{New}p_{l}\left(\boldsymbol{x_i}|\boldsymbol{\mu_l^{New}},\boldsymbol{\Sigma_l}^{New}\right)\right\}$$

Step 6

- Check for convergence of either the parameters or the log likelihood.
- If the convergence criterion is not satisfied return to step 2.

Cinvestav

References I

S. Borman, "The expectation maximization algorithm-a short tutorial," *Submitted for publication*, pp. 1–9, 2004.

J. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," *International Computer Science Institute*, vol. 4, 1998.

F. Dellaert, "The expectation maximization algorithm," tech. rep., Georgia Institute of Technology, 2002.

G. McLachlan and T. Krishnan, *The EM algorithm and extensions*, vol. 382.
John Wiley & Sons, 2007.