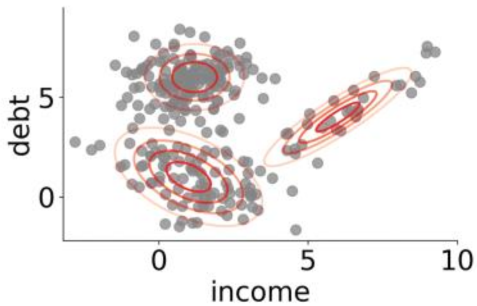
Applications and Examples

Sunday, June 24, 2018 10:11 PM

em for gmm

Applications of EM

Gaussian Mixture Model revisited



$$p(x \mid \theta) = \pi_1 \mathcal{N}(x|\mu_1, \Sigma_1) + \pi_2 \mathcal{N}(x|\mu_2, \Sigma_2) + \pi_3 \mathcal{N}(x|\mu_3, \Sigma_3)$$

$$\theta = \{\pi_1, \pi_2, \pi_3, \mu_1, \mu_2, \mu_3, \Sigma_1, \Sigma_2, \Sigma_3\}$$

Gaussain Mixture Model connection

E-step

EM: For each point compute

$$q(t_i) = p(t_i \mid x_i, \theta)$$

Gaussain Mixture Model connection

E-step

EM: For each point compute

$$q(t_i) = p(t_i \mid x_i, \theta)$$

GMM: For each point compute

$$p(t_i \mid x_i, \theta)$$

Gaussain Mixture Model connection

E-step

EM: For each point compute

$$q(t_i) = p(t_i \mid x_i, \theta)$$

GMM: For each point compute

$$p(t_i \mid x_i, \theta)$$

M-step

EM: Update parameters to maximize

$$\max_\theta \mathbb{E}_q \log p(X, T \mid \theta)$$

GMM: Update Gaussian parameters to fit points assigned to them

$$\mu_1 = \frac{\sum_i p(t_i = 1 \mid x_i, \theta)\, x_i}{\sum_i p(t_i = 1 \mid x_i, \theta)}$$

Gaussain Mixture Model connection

E-step

EM: For each point compute
$$q(t_i) = p(t_i \mid x_i, \theta)$$

GMM: For each point compute
$$p(t_i \mid x_i, \theta)$$

M-step

EM: Update parameters to maximize
$$\max_\theta \mathbb{E}_q \log p(X, T \mid \theta)$$

GMM: Update Gaussian parameters to fit points assigned to them
$$\mu_1 = \frac{\sum_i p(t_i = 1 \mid x_i, \theta)\, x_i}{\sum_i p(t_i = 1 \mid x_i, \theta)}$$

M-step derivation

$Z = \sqrt{2\pi}\,\sigma_c$

some normalization.

$$\max_\theta \sum_{i=1}^{N} \mathbb{E}_{q(t_i)} \log p(x_i, t_i \mid \theta)$$
$$= \sum_{i=1}^{N} \sum_{c=1}^{3} q(t_i = c) \log\left(\frac{1}{Z} \exp\left(-\frac{(x_i - \mu_c)^2}{2\sigma_c^2}\right) \pi_c\right)$$
$$= \sum_{i=1}^{N} \sum_{c} q(t_i = c)\left(\log\frac{\pi_c}{Z} - \frac{(x_i - \mu_c)^2}{2\sigma_c^2}\right)$$

$$\frac{\partial \dots}{\partial \mu_1} = \sum_{i=1}^{N} q(t_i = 1)\left(0 - \frac{2(x_i - \mu_1)(-1)}{2\sigma_1^2}\right) \overset{\text{let}}{=} 0 \quad \Big|\ \sigma_1^2$$

$$= \sum_{i=1}^{N} q(t_i = 1)\, x_i - \left(\sum_{i=1}^{N} q(t_i = 1) \right) \mu_1 = 0.$$

$$\mu_1 = \frac{\sum_i q(t_i = 1)\, x_i}{\sum_i q(t_i = 1)}$$

Same for μ_2, μ_3.

$$\sigma_c^2 = \frac{\sum (x_i - \mu_c)^2 q(t_i = c)}{\sum q(t_i = c)}$$

Make sure $\pi_c \geq 0$

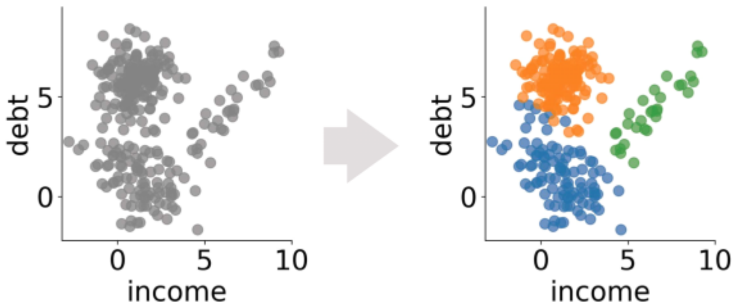
$\pi_1 + \pi_2 + \pi_3 = 1$

$$\pi_c = \frac{\sum_i q(t_i = c)}{N}$$

K-means from probabilistic perspective

k-means1

K-Means connection



K-Means

1. Randomly initialize parameters $\theta = \{\mu_1, \ldots, \mu_C\}$

K-Means

1. Randomly initialize parameters $\theta = \{\mu_1, \ldots, \mu_C\}$
2. Until convergence repeat:
   a) For each point compute closest centroid

$$c_i = \arg\min_c \|x_i - \mu_c\|^2$$

K-Means

E-M.

1. Randomly initialize parameters $\theta = \{\mu_1, \dots, \mu_C\}$
2. Until convergence repeat:
   a) For each point compute closest centroid

   $$c_i = \arg\min_c \|x_i - \mu_c\|^2$$

   b) Update centroids

   $$\mu_c = \frac{\sum_{i:c_i=c} x_i}{\#\{i : c_i = c\}}$$

K-Means from GMM perspective

From GMM to K-means:

- Fix covariances to be identical $\Sigma_c = I$
- Fix weights to be uniform $\pi_c = \dfrac{1}{\#\text{ of Guassians}}$

$$p(x_i \mid t_i = c, \theta) = \frac{1}{Z} \exp\left(-0.5\|x_i - \mu_c\|^2\right)$$

normalizing constance

center of c.

K-Means from EM perspective

E-step

$$q^{k+1} = \arg\min_{q} \mathcal{KL}\left[q(T) \parallel p(T \mid X, \theta^k)\right]$$

K-Means from EM perspective

E-step

$$q^{k+1} = \arg\min_{q \in Q} \mathcal{KL}\left[q(T) \parallel p(T \mid X, \theta^k)\right]$$

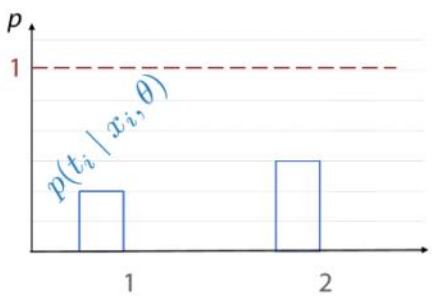
Where Q is the set of delta-functions

K-Means from EM perspective

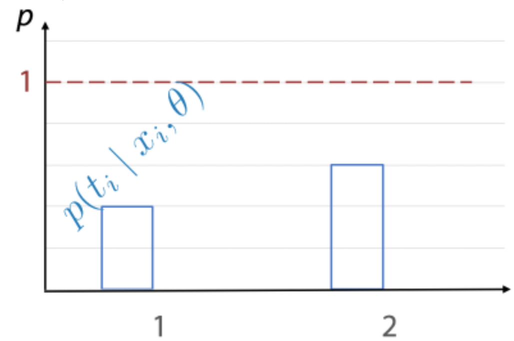
E-step

$$q^{k+1} = \arg\min_{q \in Q} \mathcal{KL}\left[q(T) \parallel p(T \mid X, \theta^k)\right]$$

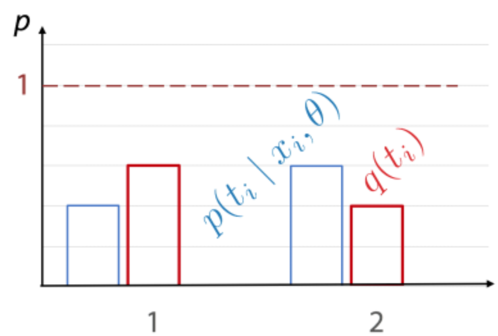
Where Q is the set of delta-functions



What is the closest approximation of the following distribution in the family of delta functions?

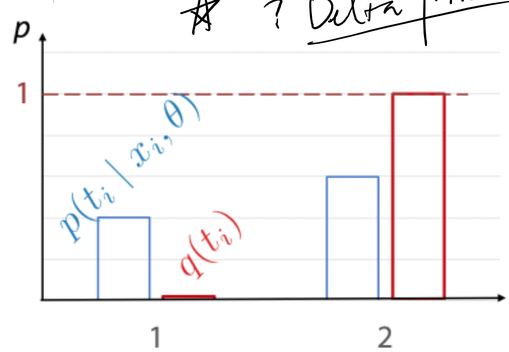


This



This

☆ ? Delta function ?

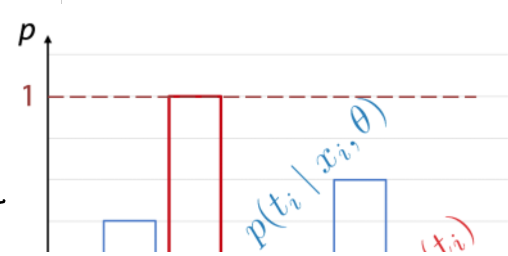


Correct

The distribution q is indeed a delta function (unlike in some other answers) and the KL divergence between this q and p is lower than the corresponding KL divergence in other answers. In this case, KL divergence equals

$$\mathcal{KL}(q \,\|\, p) = 0 \cdot \ldots + 1 \cdot \log \frac{1}{0.3} \approx 0.52$$

This



K-Means from EM perspective

E-step

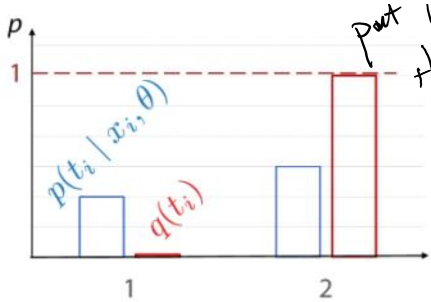
$$q^{k+1} = \arg\min_{q \in Q} \mathcal{KL}\left[q(T) \parallel p(T \mid X, \theta^k)\right]$$

Where Q is the set of delta-functions

delta function ?

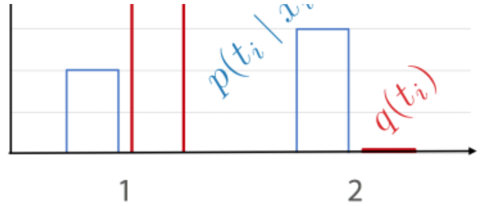
put prob. to the class higher

delta function

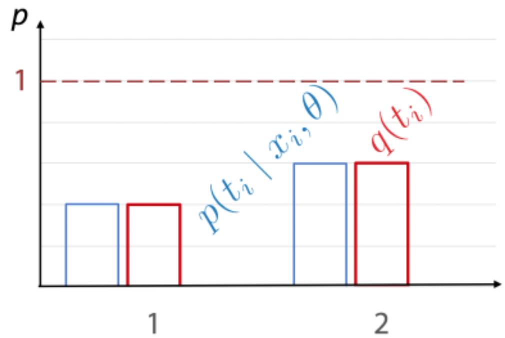


Put p

the class with higher possibilities



This



From <https://www.coursera.org/learn/bayesian-methods-in-machine-learning/lecture/qvhOF/k-means-from-probabilistic-perspective>

K-Means from EM perspective

E-step

$$q^{k+1}(t_i) = \begin{cases} 1 & \text{if } t_i = c_i \\ 0 & \text{otherwise} \end{cases}$$

K-Means from EM perspective

E-step

$$q^{k+1}(t_i) = \begin{cases} 1 & \text{if } t_i = c_i \\ 0 & \text{otherwise} \end{cases}$$

$$c_i = \arg\max_c p(t_i = c \mid x_i, \theta)$$

K-Means from EM perspective

E-step

$$q^{k+1}(t_i) = \begin{cases} 1 & \text{if } t_i = c_i \\ 0 & \text{otherwise} \end{cases}$$

$$c_i = \arg\max_c p(t_i = c \mid x_i, \theta)$$

$$p(t_i \mid x_i, \theta) = \frac{1}{Z} p(x_i \mid t_i, \theta)\, p(t_i \mid \theta)$$

K-Means from EM perspective

E-step

$$q^{k+1}(t_i) = \begin{cases} 1 & \text{if } t_i = c_i \\ 0 & \text{otherwise} \end{cases}$$

$$c_i = \arg\max_c p(t_i = c \mid x_i, \theta)$$

$$\begin{aligned} p(t_i \mid x_i, \theta) &= \frac{1}{Z} p(x_i \mid t_i, \theta)\, p(t_i \mid \theta) \\ &= \frac{1}{Z} \exp\left(-0.5\|x_i - \mu_c\|^2\right) \pi_c \end{aligned}$$

K-Means from EM perspective

q is a delta function

E-step

$$q^{k+1}(t_i) = \begin{cases} 1 & \text{if } t_i = c_i \\ 0 & \text{otherwise} \end{cases}$$

$$c_i = \arg\max_c p(t_i = c \mid x_i, \theta) = \arg\min_c \|x_i - \mu_c\|^2$$

optimal q of the E-step

$$p(t_i \mid x_i, \theta) = \frac{1}{Z} p(x_i \mid t_i, \theta)\, p(t_i \mid \theta)$$

$$= \frac{1}{Z} \exp\left(-0.5\|x_i - \mu_c\|^2\right) \pi_c$$

prior uniform. do not depend on c

can ommit

K-Means from EM perspective

E-step

$$q^{k+1}(t_i) = \begin{cases} 1 & \text{if } t_i = c_i \\ 0 & \text{otherwise} \end{cases}$$

$$c_i = \arg\min_c \|x_i - \mu_c\|^2$$

K-Means from EM perspective

E-step

$$q^{k+1}(t_i) = \begin{cases} 1 & \text{if } t_i = c_i \\ 0 & \text{otherwise} \end{cases}$$

$$c_i = \arg\min_c \|x_i - \mu_c\|^2$$

Exactly like in K-Means!

K-means M step.

To maximize

$$\max \sum_{i=1}^{N} \mathbb{E}_{q(t_i)} \log P(x_i, t_i \mid \mu)$$

~~mind~~ delta function

$$\mu_c = \frac{\sum_{i=1}^{N} q(t_i = c) \cdot x_i}{\sum_{i=1}^{N} q(t_i = c)} = \frac{\sum_{i: c_i^* = c} x_i}{\# i: c_i = c}$$

因为是hard cluster 只有1,0.

$$q(t_i) = \begin{cases} 1 & \text{if } t_i = c_i^* \\ 0 & \text{if } t_i \neq c_i^* \end{cases}$$

$\log P(X, T|\theta)$

$$\theta^{k+1} = \arg\max_{\theta} \mathbb{E}_{q^{k+1}} \log p(X, T | \theta)$$

$$\mu_c^{k+1} = \frac{\sum_{i: c_i = c} x_i}{\# \{i : c_i = c\}}$$

K-means is faster, but less flexible than GMM

k-means

K-Means from EM perspective

M-step

$$\theta^{k+1} = \arg\max_{\theta} \mathbb{E}_{q^{k+1}} \log p(X, T \mid \theta)$$

$$\mu_c^{k+1} = \frac{\sum_{i:c_i=c} x_i}{\#\{i : c_i = c\}}$$

K-Means from EM perspective

M-step

$$\theta^{k+1} = \arg\max_{\theta} \mathbb{E}_{q^{k+1}} \log p(X, T \mid \theta)$$

$$\mu_c^{k+1} = \frac{\sum_{i:c_i=c} x_i}{\#\{i : c_i = c\}}$$

Exactly like in K-Means!

K-Means from EM perspective

Summary

K-Means is actually EM for Gaussian Mixture Model, but

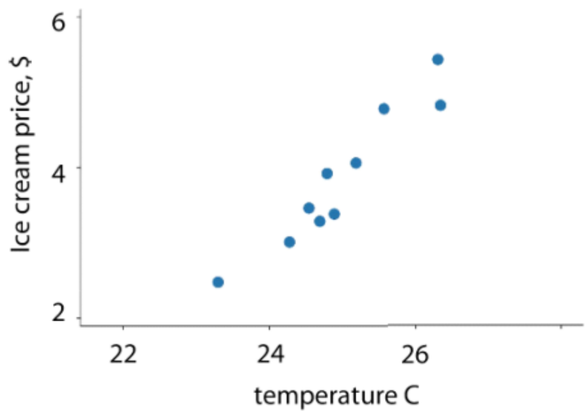
- With fixed covariance matrices $\Sigma_c = I$ (?) restricting to E step to a specific distribution.
- Simplified E-step (approximate $p(t_i \mid x_i, \theta)$ with delta function) - Connected to variation inference

Thus K-Means is faster but less flexible than GMM

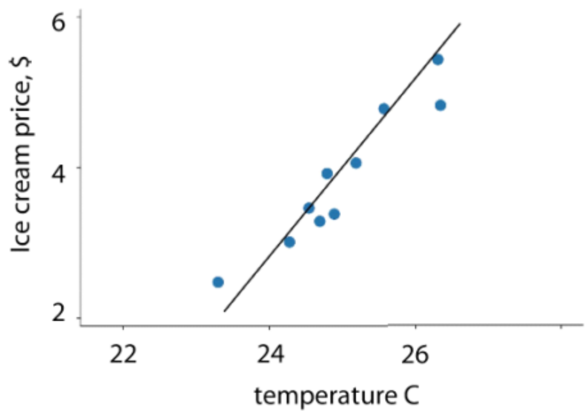
probabilistic PCA

Ice Cream conspiracy

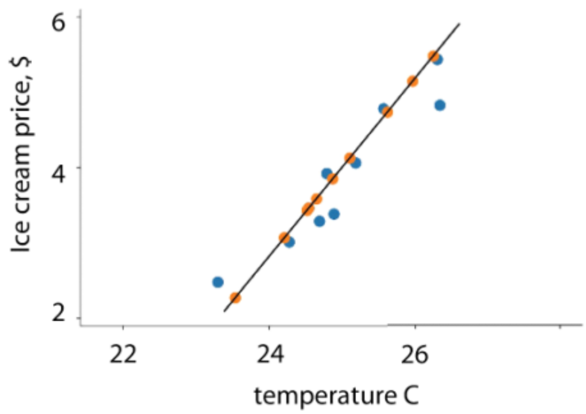
Principal Component Analysis



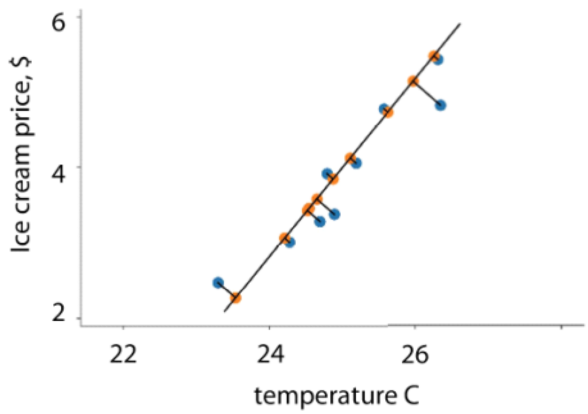
Principal Component Analysis



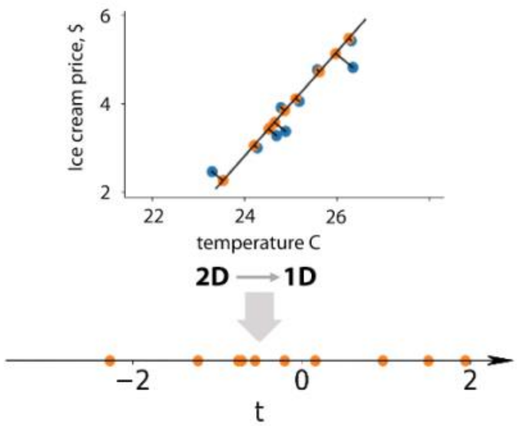
Principal Component Analysis



Principal Component Analysis



Principal Component Analysis



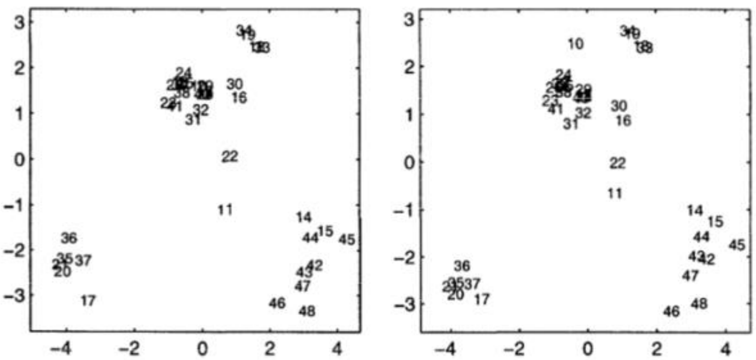
Dimensional reduction

fast

Want: formulate PCA in probabilistic terms?

why? Account for missing data

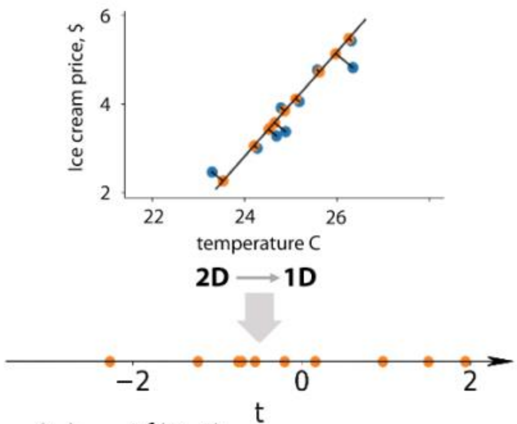
Principal Component Analysis



Projection of the Tobamovirus data by using PCA on the full data set and PPCA with 136 missing values

[source: Tipping, M. E., & Bishop, C. M. (1999). Probabilistic principal component analysis]

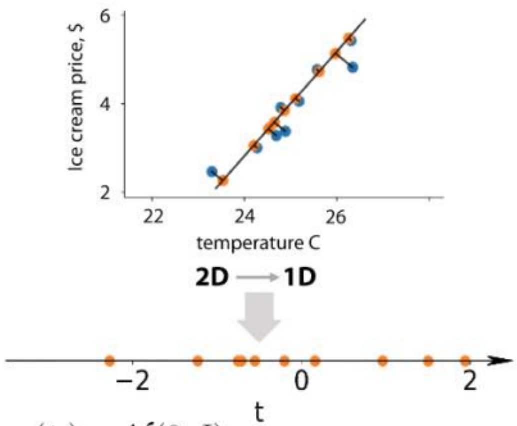
Principal Component Analysis



$$p(t_i) = \mathcal{N}(0, I)$$

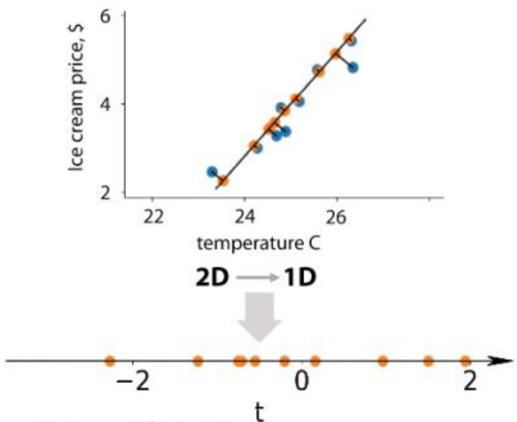
I 是啥？1？ 是1-dim的吧

Principal Component Analysis



$$p(t_i) = \mathcal{N}(0, I)$$
$$x_i = (1,1)t_i + \underbrace{(25,4)}_{\text{shift.}}$$

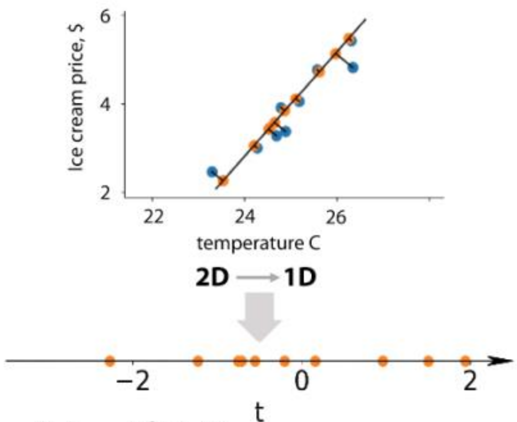
Principal Component Analysis



$$p(t_i) = \mathcal{N}(0, I)$$
$$x_i = Wt_i + b$$

orange point

Principal Component Analysis



$$p(t_i) = \mathcal{N}(0, I)$$

$$x_i = Wt_i + b + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \Sigma)$$

blue points = orange points + some noise

1-dim ⇒ 2d plus some gaussian noise

Principal Component Analysis

$$p(t_i) = \mathcal{N}(0, I)$$

$$p(x_i \mid t_i, \theta) = \mathcal{N}(Wt_i + b, \Sigma)$$

Principal Component Analysis

$$p(t_i) = \mathcal{N}(0, I)$$

$$p(x_i \mid t_i, \theta) = \mathcal{N}(W t_i + b, \Sigma)$$

$$\max_{\theta} p(X \mid \theta)$$

Principal Component Analysis

$$p(t_i) = \mathcal{N}(0, I)$$

$$p(x_i \mid t_i, \theta) = \mathcal{N}(W t_i + b, \Sigma)$$

$$\max_{\theta} p(X \mid \theta) = \prod_{i=1}^{N} p(x_i \mid \theta)$$

Principal Component Analysis

$$p(t_i) = \mathcal{N}(0, I)$$

$$p(x_i \mid t_i, \theta) = \mathcal{N}(W t_i + b, \Sigma)$$

$$\max_\theta p(X \mid \theta) = \prod_{i=1}^{N} p(x_i \mid \theta)$$

$$= \prod_i \int p(x_i \mid t_i, \theta) p(t_i) dt_i$$

=

integral. not sum. because continuous

This integral is not tractable!

So EM Algorithm will be helpful.

How is it even possible for Expectation Maximization algorithm to maximize a function ($f(\theta) = p(X \mid \theta)$) without being able to compute the value of the function, let alone the gradient $\nabla_\theta f(\theta)$?

?

- We cannot compute the value of the function (because of intractable integrals), but we can approximate it and so we can build an approximate EM algorithm which will not guarantee as the exact solution (hence no magic), but will usually work on practice.
- It's actually impossible.
- We don't need to be able to compute the value of the function to build its lower bound and to optimize this bound.

This should not be selected

It's true that we can compute *some* lower bound even if we know very little about the function (e.g. $p(X \mid \theta) \geq 0$ for any distribution p), but the lower bound will not necessarily be useful. When applying EM-algorithm, on the E-step we minimize the gap to 0 and the lower bound becomes exact at the current point ($\mathcal{L}(\theta^k, q^{k+1}) = \log p(X \mid \theta^k) = f(\theta^k)$), which means that if we can compute the lower bound $\mathcal{L}$ at any given point θ, we can compute the original function at any point as well.

Principal Component Analysis

$$p(t_i) = \mathcal{N}(0, I)$$

$$p(x_i \mid t_i, \theta) = \mathcal{N}(W t_i + b, \Sigma)$$

$$\max_\theta p(X \mid \theta) = \prod_{i=1}^{N} p(x_i \mid \theta)$$

$$= \prod_i \underbrace{\int p(x_i \mid t_i, \theta) p(t_i) dt_i}_{\text{conjugacy, } \mathcal{N}(\mu_i, \Sigma_i)} \rightarrow \text{conjugate ...}$$

The MLE of this formula is the same as PCA. waste of time? The probabilistic interpretation is still useful.

Here it is conjugate - no problem with analytical solution. No need for EM

If there's missing value — can still use EM to solve.

probabilistic PCA2

Principal Component Analysis

(Hand-waiving explanation for how EM + probabilistic PCA work)

Principal Component Analysis

E-step

$$q(t_i) = p(t_i \mid x_i, \theta)$$

Principal Component Analysis

E-step

$$q(t_i) = p(t_i \mid x_i, \theta) = \frac{p(x_i \mid t_i, \theta)p(t_i)}{Z}$$

Principal Component Analysis

E-step

$$q(t_i) = p(t_i \mid x_i, \theta) = \frac{p(x_i \mid t_i, \theta) p(t_i)}{Z}$$
$$= \mathcal{N}(\widetilde{\mu}_i, \widetilde{\Sigma}_i)$$

Principal Component Analysis

M-step

$$\max_{\theta}.\, \mathbb{E}_{q(T)} \sum_i \log p(x_i \mid t_i, \theta) p(t_i)$$

Principal Component Analysis

?

M-step

$$\max_{\theta}. \mathbb{E}_{q(T)} \sum_i \log p(x_i \mid t_i, \theta) p(t_i)$$

$$= \sum_i \mathbb{E}_{q(t_i)} \log \left(\frac{1}{Z} \exp\left(\ldots\right) \exp\left(\ldots\right) \right)$$

Principal Component Analysis

M-step

$$\max_{\theta}.\ \mathbb{E}_{q(T)} \sum_i \log p(x_i \mid t_i, \theta) p(t_i)$$

$$= \sum_i \log \frac{1}{Z}$$

$$+ \sum_i \mathbb{E}_{q(t_i)} \log \left(\exp\left(\ldots\right) \exp\left(\ldots\right) \right)$$

Principal Component Analysis

M-step

$$\max_{\theta}. \mathbb{E}_{q(T)} \sum_i \log p(x_i \mid t_i, \theta) p(t_i)$$

$$= \sum_i \log \frac{1}{Z}$$

$$+ \sum_i \mathbb{E}_{q(t_i)} \log \left(\exp\left(\ldots\right) \exp\left(-\frac{t_i^2}{2}\right)\right)$$

Principal Component Analysis

M-step

$$\max_\theta.\ \mathbb{E}_{q(T)} \sum_i \log p(x_i \mid t_i, \theta) p(t_i)$$

$$= \sum_i \log \frac{1}{Z}$$

$$+ \sum_i \mathbb{E}_{q(t_i)} \log \left(\exp \left(-\frac{(x - Wt_i - b)^2}{2\sigma^2} \right) \exp \left(-\frac{t_i^2}{2} \right) \right)$$

Principal Component Analysis

M-step

$$\max_{\theta}. \mathbb{E}_{q(T)} \sum_i \log p(x_i \mid t_i, \theta) p(t_i)$$

$$= \sum_i \log \frac{1}{Z}$$

$$+ \sum_i \mathbb{E}_{q(t_i)} \underbrace{\log \left(\exp \left(-\frac{(x - W t_i - b)^2}{2\sigma^2} \right) \exp \left(-\frac{t_i^2}{2} \right) \right)}_{a t_i^2 + c t_i + d}$$

usually hard.

can piece into normal dist's kernel

Summary

Probabilistic formulation of PCA

- Allows for missing values
- Straightforward iterative scheme for large dimensionalities
- Can do mixture of PPCA
- Hyperparameter tuning (number of components or choose between diagonal and full covariance)

Don't need linear Algebra. ^^

3. Select real-world problems which can be modeled using Gaussian Mixture Model (GMM)

- [ ] Amount of time till the next bus arrival

Un-selected is correct

- [x] Blood type distribution of people of different ethnicities

This should not be selected
Blood type is discrete variable, so it can not be modeled using Gaussian distribution.

- [x] Height distribution of people of different ethnicities

Correct
For each ethnicity we can model height using Gaussian distribution.

- [x] Rainfall measurement within 4 different seasons

Correct
For each season rainfall measurement can be modeled using Gaussian distribution.

5. Select correct statements about Probabilistic Principle Component Analysis (PPCA)

- [ ] PPCA can be computationally more efficient than naive version of its deterministic analog (PCA)

This should be selected

- [x] After training the model we can sample new data from the resulting distribution

Correct
Revise Probabilistic PCA video

- [ ] PPCA is a linear dimensionality reduction

This should be selected

- [x] PPCA can be used to visialize multidimensional data

Correct
Revise Probabilistic PCA video

4. Choose reasonable criteriums for stopping EM iterations

- [x] Parameter values stabilized (changed less than the predefines epsilon in the last iteration)

Correct

- [x] Log-likelihood lower bound stabilized (changed less than the predefines epsilon in the last iteration)

Correct

- [ ] Constraints of the original optimization problem (e.g. the prior probability weights in GMM should be non-negative and sum up to one) become satisfied

Un-selected is correct

- [ ] Log-likelihood lower bound reached the predefined constant value

Un-selected is correct