Data analysis
spatio-temporal data and hierarchical models

Bayesian Inference (Part one, Ch3)

Djemel Ziou

Learning (in materialist world)

- Goals
  - reproduction of perception, inference, behaviours, conscience, …
  - extension of the world through the creative imagination
  - enriched world = natural + artificial
- Consequences
  - more rich world, more power
  - confused world, balance of creatures
  - …
- Historical evidences
  - the progress can be slowed, but not stopped
  - the progress can be influenced by the creatures including human (knowledge, moral, interest, …).
  - philosophy, ethics, culture, and social sciences are required.

Plan

- Probability of things
- Generative learning
- Discriminative learning
- Bayesian inference
- Point estimation methods
- Approximations
- Feature selection
- References

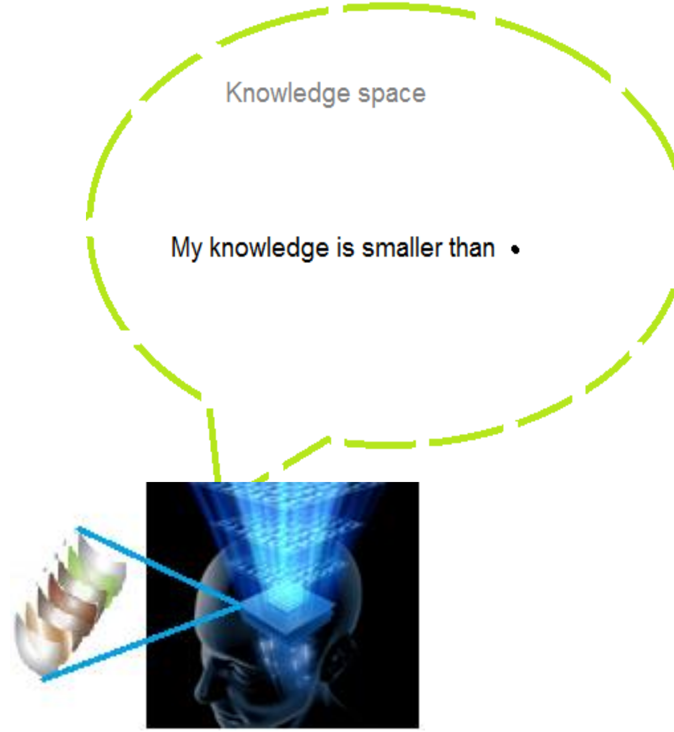
Probability of things

Probability of things

- Are we ignorant?
  - reading speed (words/second)
  - understanding (1-20 score)
  - using knowledge (1-20 score)
- To answer, I need a model

$$X = (x_1, x_2, x_3)$$

$$p(X)$$

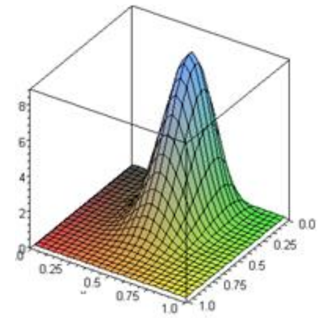


Probability of things

d-dimensional random vector $X = (x_1, \cdots, x_d)$

- Joint pdf $p(X)$
- Simplifications
  - structural (independence)

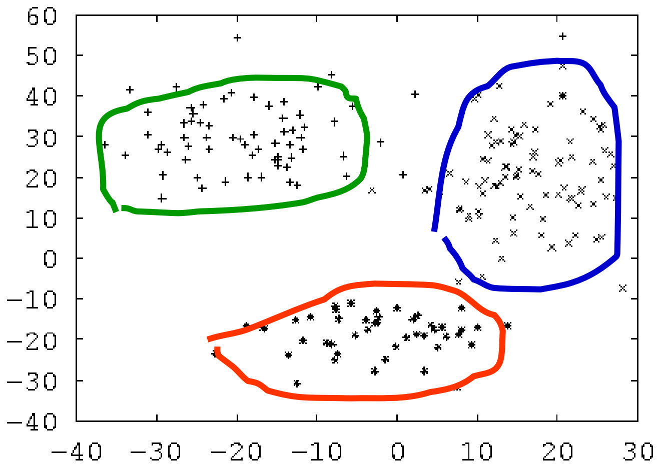
$$p(X) = \prod_i p(x_i)$$



  - parametrical

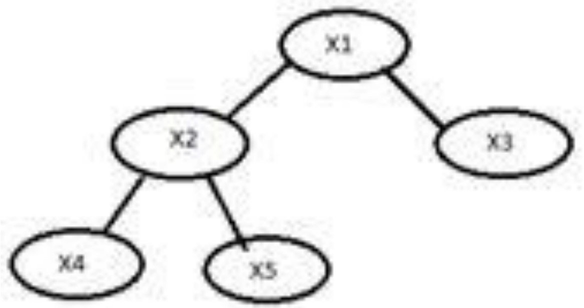
$$p(X) = \sum_k \alpha_k p_k(X)$$

$$\boldsymbol{\alpha}^t \mathbf{1} = 1$$



Probability of things

- Conditional pdf $$p(\cdots, x_i, \cdots \mid \cdots, x_{j \neq i}, \cdots)$$
- Joint/conditional $$p(X) = \prod_{i=1}^{d} p(x_i \mid x_1, \cdots, x_{j<i})$$
- Simplifications
  - structural (conditional independence)



$$p(X) = \prod_{i=1}^{d} p(x_i \mid x_{\pi_i})$$

  - parametrical (exchangeability): $(x_1, \cdots, x_d)$ are exchangeable (De Finetti)

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

The root of the Bayesian school of thought: parameter, prior on that parameter, and data IID given the parameter.

Generative learning

Generative learning

- Data: labeled or not

$$D = L \cup U;\ L = \{Y_i = (X_i, c_i)\},\quad U = \{Y_i = X_i\}$$

- Generative learning $\quad p(D) = \int p(\Theta)\, p(D \mid \Theta)\, d\Theta$
- Approximations

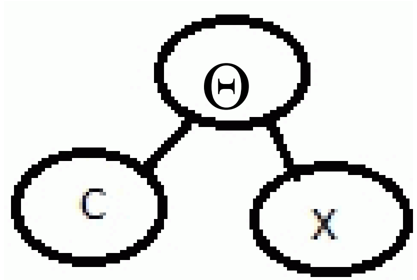
$$p(D) \approx p(D \mid \hat{\Theta}) \qquad \text{where} \quad \hat{\Theta} = \arg\max \begin{cases} p(\Theta) \prod_{y \in D} p(y \mid \Theta) & MAP \\ \prod_{y \in D} p(y \mid \Theta) & ML \end{cases}$$

- Both labeled data L and unlabeled data U can be used.

$$\hat{\Theta} = \arg\max\, p(\Theta) \prod_{y \in L} p(y \mid \Theta) \prod_{y \in U} p(y \mid \Theta)$$

- $\hat{\Theta}$ which encodes the labels of incomplete data (U) is often required (unsupervised learning).

$$p(\Theta \mid c, X)$$



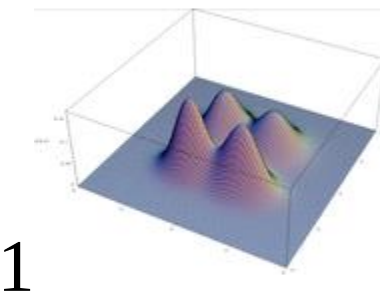
Generative learning – example

- Finite mixture
  - illiterate
  - non-graduate and cultivated
  - graduate and uncultivated
  - ...

- Applications
  - Astronomy, Biology, Economics, Engineering, Genetics, Marketing, Medicine, Psychiatry, …

Foundation of finite mixture

Let $\boldsymbol{X}_1, \ldots \boldsymbol{X}_N$ denote N random d-dimensional samples with probability density function $f(\mathbf{x}_i)$



$$f(x_i) = \sum_{j=1}^{M} \pi_j f_j(x_i)$$

where $f_j(\mathbf{x}_i)$ are densities, $\pi_i \geq 0$ and $\sum_{i=1}^{M} \pi_i = 1$

A possible interpretation

- Let Z_i be a categorical random variable taking on the values 1,… ,M with probabilities $\pi_1, \ldots, \pi_M$
- $f_j(x_i)$ (j=1,…,M) is the conditional density of $\mathbf{X}_i$ given Z_j=j
- $f(x_i)$ is a marginal pdf

$$f(x_i) = \sum_{j=1}^{M} f(x_i, z_i = j) = \sum_{j=1}^{M} p(z_i = j) f(x_i \mid z_i = j) = \sum_{j=1}^{M} \pi_j f_j(x_i)$$

Parametric formulation of mixture model

- The component densities $f_j(\mathbf{x}_i)$ are specified as $f(\mathbf{x}_i|\boldsymbol{\theta}_j)$, where $\boldsymbol{\theta}_j$ is the vector of unknown parameters
- The mixture density $f(\mathbf{x}_i)$ can then be written as

$$f(x_i \mid \Theta) = \sum_{j=1}^{M} \pi_j f(x_i \mid \theta_j)$$

where $\Theta = (\pi_1, \ldots, \pi_{M-1}, \xi^t)^t$ and $\boldsymbol{\xi} = (\theta_1, \ldots, \theta_M)$

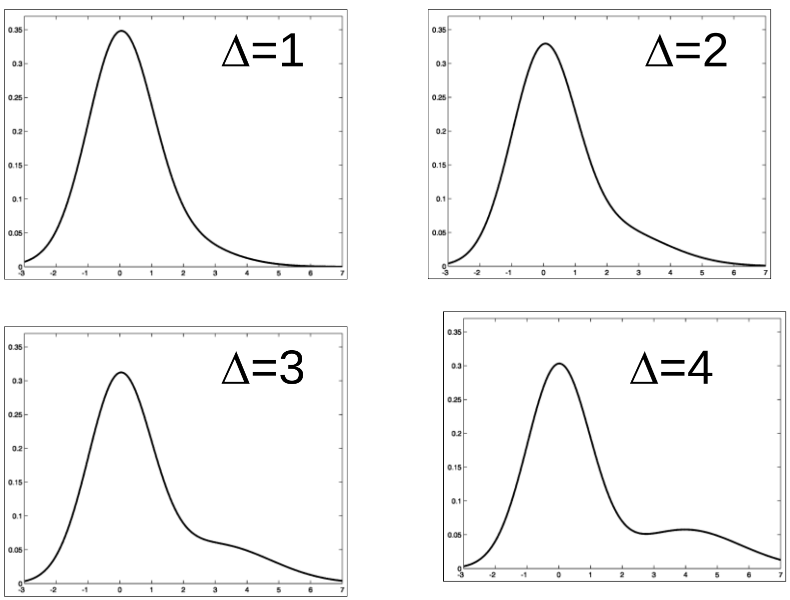
the parameters θ_j are assumed distinct.

Shapes of an univariate normal mixture

Consider $f(x_i \mid \Theta) = \pi_1 g(x_i/\mu_1, \sigma^2) + \pi_2 g(x_i/\mu_2, \sigma^2)$

where $g(x_i \mid \mu, \sigma^2) = (2\pi)^{-\frac{1}{2}} \sigma^{-1} \exp\{-\frac{1}{2}(x_i - \mu)^2 / \sigma^2\}$

μ and σ^2 are the mean and the variance.



A mixture of two univariate normal components with

$$\pi_1 = 0.75, \sigma_1 = \sigma_2 = 1$$

Generative learning

- Some properties
  - data can be incomplete
  - easy and widely used
  - the estimated model encodes information about data and labels
  - the goal is to fit the data, the target problem is not taken into account
  - MAP is more suitable when few labeled data are available

Discriminative learning

Discriminative learning

- Discriminative learning

$$p(C \mid X) = \int p(C, \Theta \mid X) d\Theta$$

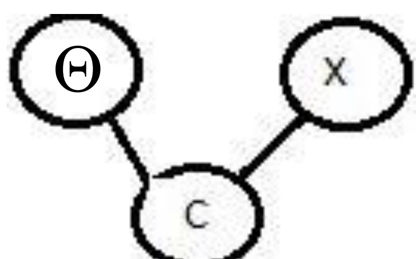
- Approximations

$$p(C \mid X) \approx p(C \mid X, \hat{\Theta}) \quad \text{where} \quad \hat{\Theta} = \arg\max \begin{cases} p(\Theta)\, p(C \mid \Theta, X) & CMAP \\ p(C \mid \Theta, X) & CML \end{cases}$$

- Unlike generative learning
  - discriminative learning cannot be used with unlabeled data

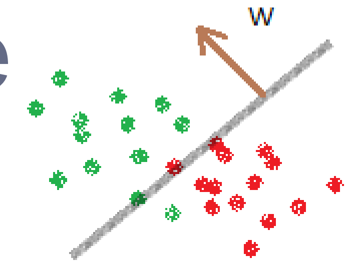
$$\hat{\Theta} = \arg\max\, p(\Theta) \prod_{y \in L} p(c \mid X, \Theta)$$

  - X provides no information about Θ; i.e., $p(X \mid \Theta) = p(X)$



$p(c \mid X, \theta)$

Discriminative learning - example



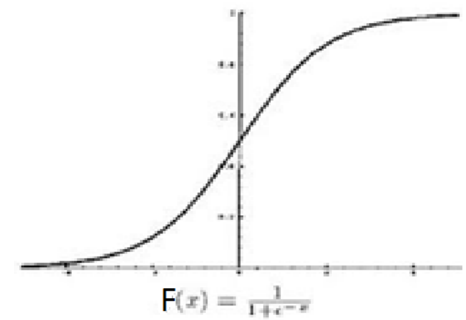
- Examples of discriminative learning models
  - logistic regression

$$p(c \mid X, W) = F((2c-1)W^t X)$$

  - probabilistic SVM approximation

$$p(c=1 \mid y, \alpha, \beta) = F(\alpha y + \beta)$$

$$y = \sum_i c_i \beta_i K(x_i, x)$$



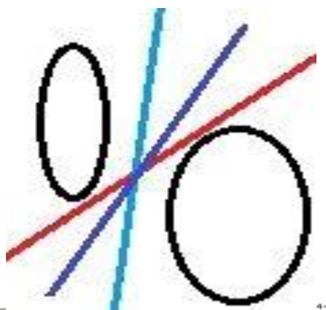
- Discriminative learning

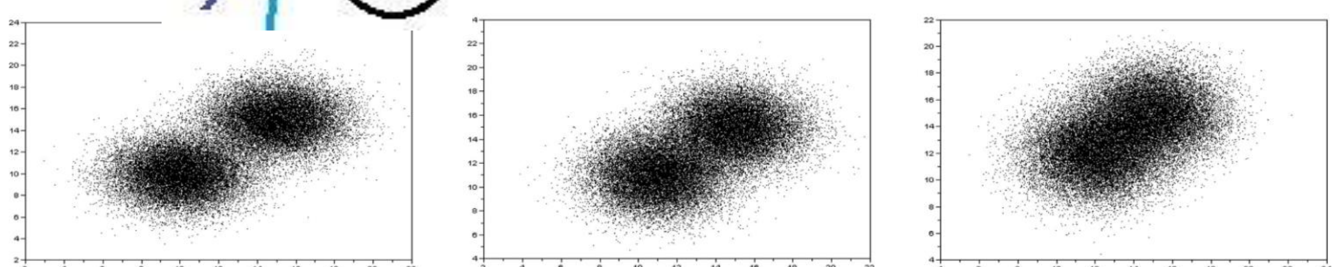
$$p(C \mid X) = \int_\Omega p(W) p(C \mid X, W) dW = \int_\Omega p(W) \prod_y F((2c-1)W^t X_i) dW$$

$$W^* \approx \begin{cases} p(W) \prod_y F((2c-1)W^t X_i) & CMAP \\ \prod_y F((2c-1)W^t X_i) & CML \end{cases}$$

Discriminative learning

- Some properties
  - data must be complete
  - discriminative learning more accurate than generative learning
  - logistic regression outperform SVM and its variants
  - model encodes information only about labels
  - CMAP is suitable when few labeled data are available
  - separability problem when CML is used





(a) (b) (c)

Bayesian inference

Goal

- Data are generated from unknown distribution
- By analyzing the data, we would like to learn about the distribution, predict some future data, and make decision
- The common methodology is to assume that the distribution is known except the value of some of its parameters (parametric methods).
- In this case, one issue is to estimate the parameters (learning).
- Bayesian inference involves prior, posterior, and likelihood distributions.

Bayesian data analysis

- Prediction $p(x \mid D)$
- Decision making $r(a \mid x) = \sum_i l(a, w_i)\, p(w_i \mid x)$
- Estimation
  - for fundamentalists, the posterior is the end result and no estimation is allowed
  - in practice, the posterior is summarized by using an estimate

$$\min_{\hat{\Theta}} \int_{\Omega} l(\hat{\Theta}, \Theta)\, p(\Theta \mid D)\, d\Theta$$

  - Examples

$$l(\hat{\Theta}, \Theta) = 1 - \delta(\hat{\Theta} - \Theta) \quad \Rightarrow \quad \hat{\Theta} = \arg\max_{\theta} p(\Theta \mid D)$$

$$l(\hat{\Theta}, \Theta) = (\hat{\Theta} - \Theta)^2 \quad \Rightarrow \quad \hat{\Theta} = \int_{\Theta} \Theta\, p(\Theta \mid D)\, d\Theta$$

  - Exercise: evaluate the last estimator in the case of a normal pdf with known variance. You consider that the mean is sampled from normal prior.

Bayesian inference - foundation

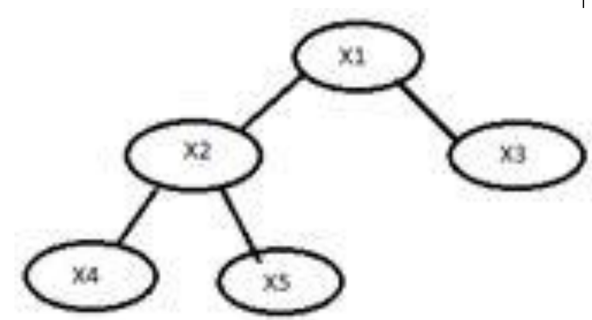
- Hidden (missed, latent) variables helps in
  - labelling samples, objects, classes, patterns, …
  - modeling non observed phenomena
  - combining variables in order to reduce dimension
- They are unobserved (non measurable)
- They can be introduced in a model by marginalization
- Example, the marginalized (integrated) likelihood

$$p(D \mid \Theta) = \int_{\Omega} p(z \mid \Theta)\, p(D \mid z, \Theta)\, dz$$

$$p(z \mid D, \Theta) = p(D \mid z, \Theta)\, p(z \mid \Theta) \Big/ \int_{\Omega} p(D \mid z, \Theta)\, p(z \mid \Theta)\, dz$$

Bayesian inference - foundation

- Graphical model G represents dependencies between variables
- G=(V,E) be a directed acyclic graph



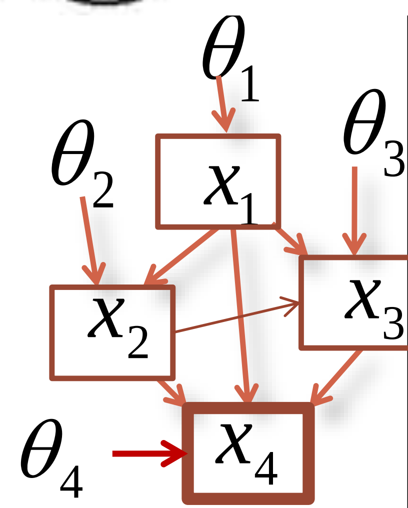
$$p(x_1, \ldots, x_d) = \prod_{i=1}^{d} p(x_i \mid x_{\pi_i})$$

Bayesian network

- Example

$$p(x_1, x_2, x_3, x_4 \mid \Theta) = p(x_1 \mid \theta_1)\, p(x_2 \mid x_1, \theta_2)$$

$$p(x_3 \mid x_1, x_2, \theta_3)\, p(x_4 \mid x_1, x_2, x_3, \theta_4)$$



- Undirected graph : See Markov random field.

Bayesian inference - foundation

- In order to estimate Θ by using marginalized likelihood, the challenge is to compute the integral

$$p(D \mid \Theta) = \int_z p(z \mid \Theta) p(D \mid z, \Theta) dz$$

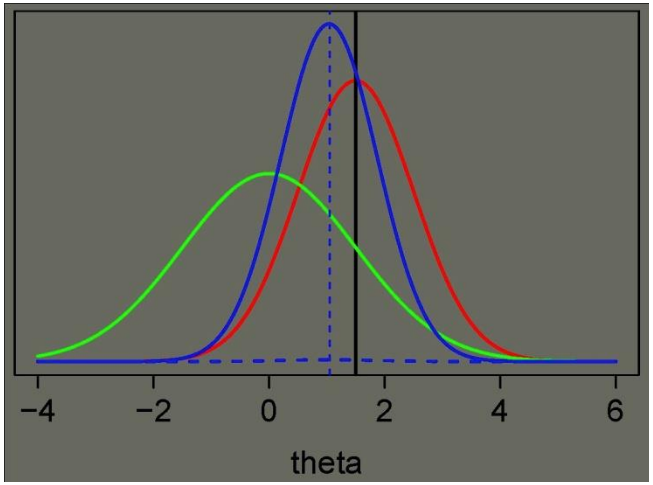
- Example $p(z \mid \alpha) \propto e^{-z^2/\alpha}$ $\quad p(D \mid z, \alpha) \propto \prod_{i=1}^{N} e^{-(z x_i)^4}$
- Several methods exist
  - simulation
  - Laplace approximation
  - ...
  - variational approximation

Prior and posterior pdfs

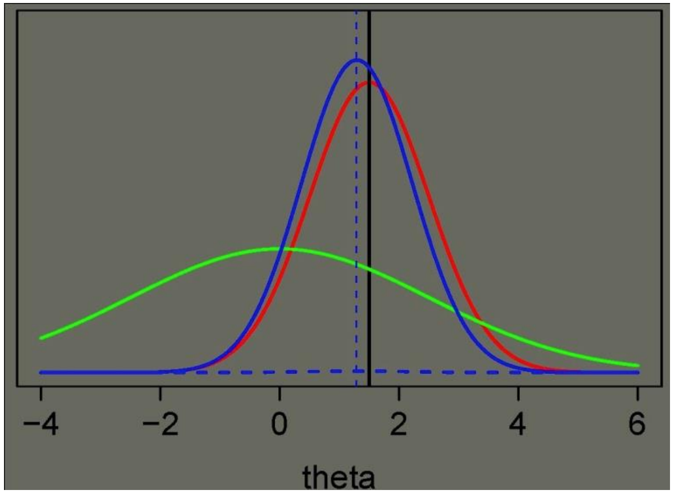
- Let us consider the space of parameters of a pdf. Prior distribution indicates the region in which lies the parameters.
- Prior pdf can be determined from information about frequencies (e.g., previous experimentation) or subjective information.
- They can be a pdf or not.
- Improper prior pdf $f(\theta)$ are such that $\int_\Omega f(\theta)d\theta=+\infty$. In this case, the posterior is not necessarily proper. This is used when we would like to give more importance to the data.
  - $\theta^{-1}(1-\theta)^{-1}$ where $\theta\in[0,1]$ and $f(\sigma)=\sigma^{-1}$ are improper priors
- Non informative prior is often determined by assigning equal probability to all values (principle of indifference). Unlike improper prior, it is a pdf and therefore the posterior.
  - $f(\theta)$=constant is non normative. In this case, the maximum likelihood and the MAP are the same.

Posterior and prior

- Conjugate prior is a pdf in the same family as the posterior. This choice of prior may allows to make the integral in the denominator tractable.
  - example, $g(x|\mu)$ a Gaussian, $g(\mu|\mu_0,\sigma_0)$ a Gaussian, and σ known



Prior N(0,0.5), Likelihood, Posterior
Prior not very informative



Prior N(0,2.5), Likelihood, Posterior
Informative prior

Estimation methods

- The posterior is used for the estimation of parameters
  - Full-Bayesian: use of loss function

$$l(\hat{\Theta}, \Theta) = (\hat{\Theta} - \Theta)^2 \quad \Rightarrow \quad \hat{\Theta} = \int_{\theta} \Theta p(\Theta \mid D) d\Theta$$

    not easy to determine the suitable loss function

  - Bayesian-point: maximum likelihood, maximum *a posteriori*, simulation…
  - approximations are sometimes required: Laplace approximation, variational approximation…

Point estimation methods

Point estimation of parameters

- d-dimensional random vector $x = (x_1, \cdots, x_d)^t$
- Observations $D = \{x_1, \cdots, x_N\}$
- Model estimation
  - maximum likelihood
  - expectation maximization
  - maximum *a posteriori*
  - maximum entropy, moment, simulation…
- Model selection

Maximum likelihood

- Indented for the model estimation (not clustering).
- There exist several formulations. We will study the most common.
- Likelihood, under the iid assumption

$$p(D \mid \Theta) = \prod_{n=1}^{N} p(x_n \mid \Theta)$$

- Maximum likelihood estimator $\hat{\Theta}$

$$L(\hat{\Theta}) = \max_{\Theta} p(D \mid \Theta)$$

under constraints (if any).

Maximum likelihood

- Example-Mixture estimation
  - likelihood, under the iid assumption

$$p(D|\Theta) = \prod_{n=1}^{N} p(x_n|\Theta) = \prod_{n=1}^{N} \sum_{j=1}^{K} \pi_j f(x_n|\theta_j)$$

  - maximum likelihood

$$\max_{\Theta} L(\Theta) = \prod_{n=1}^{N} \sum_{j=1}^{K} \pi_j f(x_n|\theta_j)$$

constraints $\pi_j \geq 0$ and $\sum_{j=1}^{K} \pi_j = 1$

  - optimization problem

$$\min_{\Theta} - \sum_{n=1}^{N} \ln\left(\sum_{j=1}^{K} \pi_j f(x_n|\theta_j)\right) + \lambda\left(1 - \sum_{j=1}^{K} \pi_j\right)$$

Maximum likelihood

- first order condition

$$\forall k = 1,\dots,K \quad -\sum_{n=1}^{N} \frac{\dfrac{\partial f(x_n \mid \theta_k)}{\partial \theta_k}}{\sum_{j=1}^{K} \pi_j f(x_n \mid \theta_j)} = 0$$

$$\forall k = 1,\dots,K \quad -\sum_{n=1}^{N} \frac{f(x_n \mid \theta_k)}{\sum_{j=1}^{K} \pi_j f(x_n \mid \theta_j)} - \lambda = 0$$

$$\sum_{j=1}^{K} \pi_j - 1 = 0$$

- solve the (often nonlinear) system of equations

Maximum likelihood

- iterative algorithm
  - input $D, K, \theta_1^0, \ldots, \theta_K^0, \pi_1^0, \ldots, \pi_{K-1}^0$
  - output $\hat{\theta}_1, \ldots, \hat{\theta}_K, \hat{\pi}_1, \ldots, \hat{\pi}_K$
  - r=0
  - do

    r = r+1

    for k=1 to K

    $$\theta_k^{(r)} = \ldots$$

    $$\pi_k^{(r)} = \ldots$$

    end for

    until convergence
- for convergence, the likelihood function can be used
- for the initial values: see k-means, fuzzy c-means, ….
- check the second order condition

Maximum likelihood

To illustrate, let us consider a mixture of two univariate Gaussian with common variance

$$f(x \mid \Theta) = \pi_1 g(x \mid \mu_1, \sigma^2) + \pi_2 g(x \mid \mu_2, \sigma^2) \quad \text{where}$$

$$g(x \mid \mu, \sigma^2) = (2\pi)^{-\frac{1}{2}} \sigma^{-1} \exp(-\tfrac{1}{2}(x-\mu)^2 / \sigma^2) \text{ and } \Theta = (\mu_1, \mu_2, \sigma)^t$$

the optimization problem $\min_{\Theta} - \sum_{n=1}^{N} \ln(\sum_{j=1}^{2} \pi_j g(x_n \mid \mu_j, \sigma)) + \lambda(1 - \sum_{j=1}^{2} \pi_j)$

the first order condition

$$\forall k = 1,2 \quad -\sum_{n=1}^{N} \frac{\frac{\partial g(x_n \mid \theta_k)}{\partial \mu_k}}{\sum_{j=1}^{2} \pi_j g(x_n \mid \theta_j)} = 0 \qquad \forall k = 1,2 \quad -\sum_{n=1}^{N} \frac{g(x_n \mid \theta_k)}{\sum_{j=1}^{2} \pi_j g(x_n \mid \theta_j)} - \lambda = 0$$

$$\sum_{j=1}^{2} \pi_j - 1 = 0$$

Maximum likelihood

N=50

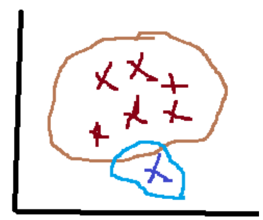
Estimation of mixing proportions

Iteration r	$\pi_1^{(r)}$	$\log L(\pi_1^{(r)})$
0	**0.50000**	**-91.87811**
1	**0.68421**	**-85.55353**
2	**0.70304**	**-85.09035**
3	**0.71792**	**-84.81398**
6	**0.74218**	**-84.60978**
7	**0.74615**	**-84.58562**
	0.50000	**-91.87811**
27	**0.68421**	**-85.55353**

Ground truth $\mu_1 = 0$ $\mu_2 = 2$ $\sigma^2 = 1$ $\pi_1 = 0.8$ $\pi_2 = 0.2$

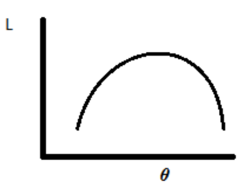
Maximum likelihood

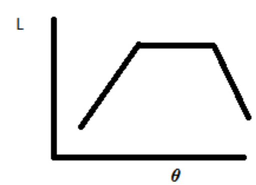
- Some properties
  - easy to derive and to use for supervised learning
  - good estimator; may be consistent, unbiased, efficient
  - invariant; if $t = r(\Theta)$ then $\hat{t} = r(\hat{\Theta})$
  - limited to iid
  - no prior knowledge
  - may not exist



  - point estimator (poor uncertainty)

  - numerical issues; flat maximum, several maxima, degenerate maximum





Expectation maximization

- EM intended for clustering.
- The main idea is data can be viewed as being complete $(\mathbf{x}_1, \mathbf{z}_1)$,…, $(\mathbf{x}_N, \mathbf{z}_N)$, where $z_i = (\underbrace{0,0,\ldots,1,0,\ldots,0}_{j^{th}})^t$
- We wish to infer the $\mathbf{z}_i$ on the basis of the feature data $\mathbf{x}_i$.
  - estimation of the model $\hat{\Theta}$
  - for each $\mathbf{x}_j$, $f(\hat{\theta}_1 \mid \mathbf{x}_j)$,…, $f(\hat{\theta}_K \mid \mathbf{x}_j)$ are the estimated posterior probabilities that this observation belongs to the 1st, 2nd,…, and Kth component. Indeed, the estimated component-label vector $\mathbf{z}_i$ is $\hat{\mathbf{z}}_i$ where $\hat{\mathbf{z}}_{ij} = (\hat{\mathbf{z}}_{ij})$

$$\hat{z}_{ij} = \begin{cases} 1 & if \;\; j = \arg\max_h f(\hat{\theta}_h \mid x_i) \\ 0 & elsewhere \end{cases}$$

Expectation maximization

- Algorithm
  - input $D, K, \theta_1^0, \ldots, \theta_K^0, \pi_1^0, \ldots, \pi_{K-1}^0$
  - output $\hat{\theta}_1, \ldots, \hat{\theta}_K, \hat{\pi}_1, \ldots, \hat{\pi}_K$
  - do
    - E-Step, estimate the posterior

      for each j and i $\tau_{ji} = f(\theta_j^{(r-1)} \mid x_i) = \pi_j^{(r-1)} f(x_i \mid \theta_j^{(r-1)}) / f(x_i \mid \Theta^{(r-1)})$

    - M-Step, fond Θ by maximizing

$$\sum_{n=1}^{N} \sum_{j=1}^{K} \tau_{nj} \ln(\pi_j^{(r-1)} f(x_i \mid \theta_j^{(r-1)}))$$

    Until convergence

- clustering (hard clustering if needed)

$$\hat{z}_{ij} = \begin{cases} 1 & if \;\; j = \arg\max_h f(\hat{\theta}_h \mid x_i) \\ 0 & elsewhere \end{cases}$$

Expectation maximization

- Example, let us consider the normal mixture

$$g(x_i \mid \Theta) = \sum_{j=1}^{K} \pi_j g(x_i \mid \theta_j)$$

- E-step

$$\tau_{ij}^{(r)} = g(\theta_j^{(r)} \mid x_i) = \pi_i^{(r)} g(x_i \mid \theta_j^{(r)}) / g(x_i \mid \Theta^{(r)})$$

- M-Step

$$\pi_j^{(r+1)} = \sum_{i=1}^{N} \tau_{ji}^{(r)} / N \qquad (\mathrm{j} = 1, \ldots, \mathrm{K})$$

$$\mu_j^{(r+1)} = \sum_{i=1}^{N} \tau_{ji}^{(r)} x_i / \sum_{i=1}^{N} \tau_{ji}^{(r)} \qquad (\mathrm{j} = 1, \ldots, \mathrm{K})$$

$$\sigma_j^{2(r+1)} = \sum_{i=1}^{N} \tau_{ji}^{(r)} (x_i - \mu_j^{(r+1)})^2 / \sum_{i=1}^{N} \tau_{ji}^{(r)} \qquad (\mathrm{j} = 1, \ldots, \mathrm{K})$$

Expectation maximization

- Some properties
  - converges to the maximum likelihood
  - easy and allows clustering (incomplete data) and supervised learning
  - the K dimensional problem is split to 1D problems
  - share properties with the maximum likelihood.

Maximum a posteriori

- The MAP $\hat{\Theta}_{MAP} = \arg\max_{\Theta} p(\Theta \mid D)$
  - the integral is often intractable

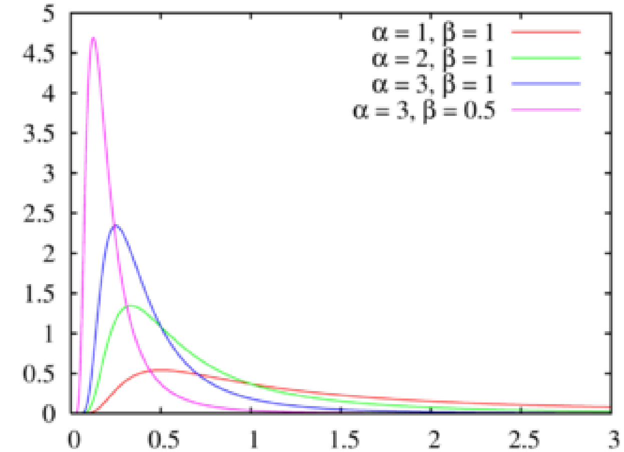
$$p(\Theta \mid D) = p(D \mid \Theta)\, p(\Theta) / \int_{\Omega} p(D \mid \Theta)\, p(\Theta) d\Theta$$

  - simplification $\hat{\Theta}_{MAP} = \arg\max_{\Theta} p(D \mid \Theta)\, p(\Theta)$

Maximum a posteriori

- Example
  - likelihood

$$p(D\mid\Theta) \propto \prod_{j=1}^{N} e^{-\frac{(x_j-\mu)^2}{2\sigma^2}}$$



  - prior

$$p(\Theta) = p(\mu)\,p(\sigma) \propto e^{-\frac{(\mu-\eta)^2}{2\lambda^2}} \frac{\beta^\alpha}{\Gamma(\alpha)} \sigma^{-\alpha-1} e^{-\beta/\sigma}$$

  - posterior

$$\max_{\mu,\sigma} p(\Theta\mid D) \propto \prod_{j=1}^{N} e^{-\frac{(x_j-\mu)^2}{2\sigma^2}} e^{-\frac{(\mu-\eta)^2}{2\lambda^2}} \frac{\beta^\alpha}{\Gamma(\alpha)} \sigma^{-\alpha-1} e^{-\beta/\sigma}$$

  - hyperparameters $\eta, \lambda, \beta, \alpha$

Maximum a posteriori

- Some properties
  - easy to derive and to use for supervised learning
  - another interpretation $\hat{\Theta}_{MAP} = \arg\max_{\Theta} p(D,\Theta)$
  - good estimator; may be consistent, unbiased, efficient
  - looks like the maximum likelihood when N tends to infinity
  - unlike the maximum likelihood, it avoids the overfitting (overfitting: error decrease in learning phase and increase in the test phase)
  - non invariant
  - limited to iid
  - may not exist (see maximum likelihood)
  - point estimator (see maximum likelihood)
  - numerical issues; flat maximum, several maxima, degenerate maximum.

Model selection

- Find $\Theta \in \{\Theta_1, \Theta_2, \ldots\}$
- Example of methods
  - Akaike's Information Criterion (AIC) selects the model that minimizes

$$-2\log L(\hat{\Theta}) + 2d$$

    where d is equal to the number of parameters in the model.
  - the Bayesian information criterion (BIC) of Schwarz (1978) is given by

$$-2\log L(\hat{\Theta}) + d\log N$$

  - many other criteria MDL, MML, …

Model selection

- Algorithm
  - input D, $\{\Theta_1, \Theta_2, \ldots\}$
  - output $\hat{\Theta}$
  - min=+∞
  - for $\Theta \in \{\Theta_1, \Theta_2, \ldots\}$

$$\textit{if } (v = -2\log L(\Theta) + \textit{complexity term}) < \min) \textit{ then}$$

$$\hat{\Theta} = \Theta; \min = v; \textit{ EndIf}$$

  Endfor
- The set $\{\Theta_1, \Theta_2, \ldots\}$ can be estimated using the maximum likelihood, EM…

Model selection

- Example – unsupervised learning of a mixture of pdfs
  - input D
  - output $\hat{\Theta}$
  - min= $+\infty$
  - for j=1..K_{max}
    - estimate Θ for a mixture of j components

$$if\ (v = -2\log L(\Theta) + complexity\ term) < \min)\ then$$

$$\hat{\Theta} = \Theta; \min = v;\ EndIf$$

Endfor

- The set $\{\Theta_1, \Theta_2, \ldots\}$ can be estimated using the maximum likelihood, EM…

Model selection

- Some properties
  - intended for unsupervised learning
  - inherits the ML and MAP properties

References

- Model selection
  - N. Bouguila and D. Ziou. High-Dimensional Unsupervised Selection and Estimation of a Finite Generalized Dirichlet Mixture Model Based on Minimum Message Length. IEEE Transactions on Pattern Analysis and Machine Intelligence 29, pp. 1716 - 1731 2007.
- Maximum likelihood
  - In Jae Myung, Tutorial on maximum likelihood estimation. Journal of Mathematical Psychology 47, pp. 90–100, 2003.