Short course

A vademecum of statistical pattern recognition and machine learning

Density estimation

Massimo Piccardi
University of Technology, Sydney, Australia

Agenda

- Density estimation
- Likelihood function
- Maximum-likelihood density estimation
- The Expectation-Maximisation (EM) approach
- EM for Gaussian mixture models
- Maximum-a-posteriori density estimation
- Non-parametric density estimation: Kernel Density Estimation
- Example papers

Density estimation

- Accurate **modelling of pdfs and probabilities from sets of samples** is a fundamental task in pattern recognition
- This task is now universally referred to as model's **learning**
- A most obvious use of a pdf trained from data from one class is as class-conditional likelihood, p(x | class), in Bayesian classification: p(class | x) $\propto$ p(x | class) $\cdot$ p(class); yet, likelihoods are used also in other contexts
- (probability) density (function) estimation
  - Parametric: Gaussians, Gaussian mixtures etc
  - Non-parametric: histogram, k-nearest neighbours, KDE, mean-shift etc
- Gaussian, Gaussian mixtures and KDE in the following

Predictive distribution

- We are given a set of samples, X = {x_i}, i=1..N, and we are interested in computing the probability of a future sample, x, based on X: **p(x | X)**. This is often called the *predictive distribution*
- We introduce another random variable representing a set of parameters, θ, and we write:

$$p(x \mid X) = \int_{\theta} p(x, \theta \mid X)\, d\theta =$$

$$= \int_{\theta} p(x \mid \theta, X)\, p(\theta \mid X)\, d\theta$$

- The above holds thanks to Bayes' theorem and marginalisation

Predictive distribution

- Now we introduce an approximation: that the set of parameters, θ, can "summarise" X to the purpose of the prediction:

$$p(x \mid X) = \\ = \int_{\theta} p(x \mid \theta, X) p(\theta \mid X) d\theta \approx \int_{\theta} p(x \mid \theta) p(\theta \mid X) d\theta$$

- In more technical terms, we assume that x is independent of X when conditioned on θ. This is called a *parametric approach*

Predictive distribution

- As a further assumption, we assume that all the volume of pdf p(θ|X) is concentrated in one point, θ*:

$$p(x \mid X) \approx \\ \approx \int_{\theta} p(x \mid \theta) p(\theta \mid X) d\theta \approx p(x \mid \theta^*)$$

the above is called a *point estimation* for the predictive distribution

Predictive distribution

- If we further assume to choose θ* such that:

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\ p(\theta \mid X)$$

the above is called **maximum-a-posteriori estimation (MAPE)**

- By expressing p(θ |X) as ∝ p(X| θ) p(θ) and assuming p(θ) uniform, we can choose θ* as:

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\ p(X \mid \theta)$$

the above is called **maximum-likelihood estimation (MLE)**

Predictive distribution

- This brief introduction shows three possible style of predictions:
  - computing p(x|X) by marginalising θ (full Bayesian treatment of the parameters)
  - with a point-estimate approximation, computing p(x|θ*) by choosing θ* that maximises p(θ|X) (MAP estimation)
  - with a further approximation, computing p(x|θ*) by choosing θ* that maximises p(X|θ) (ML estimation)
- We have made the assumption that p(x|θ,X) ≈ p(x|θ). This assumption does not hold for the so-called *nonparametric* methods, where data X need to be retained for prediction

Taxonomy

	parametric	**nonparametric**
Bayesian	$p(x \mid X) \approx \int_{\theta} p(x \mid \theta) p(\theta \mid X) d\theta$	$p(x \mid X) = \int_{\theta} p(x \mid \theta, X) p(\theta \mid X) d\theta$
Point estimate	$p(x \mid X) \approx p(x \mid \theta^*)$	$p(x \mid X) \approx p(x \mid \theta^*, X)$

MAPE: $\theta^* = \underset{\theta}{\operatorname{argmax}}\ p(\theta \mid X)$

MLE: $\theta^* = \underset{\theta}{\operatorname{argmax}}\ p(X \mid \theta)$

Maximum-likelihood density estimation

- Let us have a set of samples, X = {x_i}, i=1..N, and consider their joint probability, p(X) = p(x_1, ... x_i, ... x_N)
- We assume that this joint probability is parametric in some parameters, θ, and thus noted p(X|θ)
- This dependence is a functional dependence and the corresponding function is known as the *likelihood function*, noted as L(θ|X) for clarity
- In **maximum-likelihood density estimation (MLE)**, our goal is to find θ such that:

$$\theta = \arg\max_{\theta} \left(L(\theta \mid X) \equiv p(X \mid \theta) \right)$$

Likelihood function

- If the samples are all generated from the same distribution and independently of one another (so called *independently and identically distributed* (*i.i.d.*) samples), the joint probability of the entire set, X, is then given by:

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

- Density estimation is performed by choosing an appropriate pdf model (for example, Gaussian) and **fitting its parameters**, θ, to the set of samples so as to maximise the likelihood function
- Important caveat! For some models, the likelihood function has multiple local maxima and finding the global one may prove hard

Log-likelihood

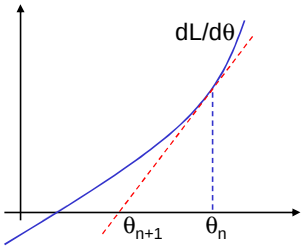
- The log-likelihood, $LL(\theta|X) = \ln L(\theta|X)$, is often used instead of L for these main reasons:
  - log (or ln) is a monotically increasing function of its argument: maxima of the argument are maxima also of its log (this is a “non-disadvantage”)
  - $\log \prod = \sum \log$: removes the product operator which is often harder to deal with in maximisations
  - for distributions of the exponential family, the log cancels with the exp
  - working in a logarithmic scale reduces the occurrence of numerical underflow/overflow during the evaluation of L (of course, remapping numerical resolution); example:

$$\prod_{1000} 0.1 = 10^{-1000} \quad \leftrightarrow \quad \sum_{1000} -1 = -1000$$

- The base of the logarithm does not really matter for the maximisation; it often is *e* (the Euler’s number)

Likelihood maximisation

- Find parameters θ maximising $L(\theta)$ or, equivalently, $LL(\theta)$
- It is possible to undertake direct maximisation by differentiation: compute $\nabla L = 0$ and solve for θ (if θ is m-dimensional, m partial derivatives)
- If closed-form solutions are not possible, iterative methods can be used instead (e.g. Newton-Raphson, requiring the m x m Hessian, or approximated methods)



θ_{n+1} is a better approximation than θ_n for the zero of $dL/d\theta$

ML for a discrete distribution

- Easy case allowing a closed-form solution
- Let us assume a discrete random variable, x, with L possible outcomes
- The parameters of this distribution are the L probability values, π_l, constrained as (*simplex* constraint):

$$\begin{cases} (1 \geq)\, \pi_l \geq 0, \quad l = 1 \ldots L \\ \sum_{l=1}^{L} \pi_l = 1 \end{cases}$$

- The probability mass function can be noted as $p(x \mid \pi)$, where $\pi = [\pi_1 \ldots \pi_L]$

Likelihood function for the discrete distribution

- Assuming a set of N independent samples, X = {x_i}, i = 1 … N, all from the same distribution, p(x|π), its likelihood function is:

$$L(\pi \mid X) = p(x_1, \ldots x_N \mid \pi) = \prod_{i=1}^{N} p(x_i \mid \pi)$$

- Let us move to logarithmic scale (log-likelihood):

$$LL(\pi \mid X) = \ln p(x_1, \ldots x_N \mid \pi) = \sum_{i=1}^{N} \ln p(x_i \mid \pi)$$

Likelihood function for the discrete distribution

- The probability for every sample x that has value l is π_l by definition. We can therefore re-write the log-likelihood function in terms of the number of samples that have the same value:

$$I(x_i, l) = \begin{cases} 1 & if\ x_i = l \\ 0 & otherwise \end{cases} \qquad n_l = \sum_{i=1}^{N} I(x_i, l)$$

$$\rightarrow LL(\pi \mid X) = n_1 \ln \pi_1 + n_2 \ln \pi_2 + \ldots = \sum_{l=1}^{L} n_l \ln \pi_l$$

Lagrangian equation

- We now need to maximise LL(π | X) *subject to the simplex constraint*. This is a case of constrained optimisation and we can solve it by maximising a corresponding *Lagrangian equation*:

$$Lag(\pi,\lambda)=\sum_{l=1}^{L} n_l \ln \pi_l + \lambda\left(\sum_{l=1}^{L}\pi_l - 1\right)$$

- Constraints $\pi_l \geq 0$, l = 1…L, are implied by the argument of the logarithm
- λ is a Lagrangian multiplier

Maximum-likelihood parameters

- Let us differentiate the Lagrangian in all π_l and equate to 0, and eliminate λ using the results and the constraint:

$$\frac{\partial}{\partial \pi_l}\left[\sum_{l=1}^{L} n_l \ln \pi_l + \lambda\left(\sum_{l=1}^{L}\pi_l - 1\right)\right] = \frac{n_l}{\pi_l} + \lambda = 0 \quad \rightarrow \lambda\pi_l = -n_l$$

$$!\rightarrow \lambda\sum_{l=1}^{L}\pi_l = -\sum_{l=1}^{L} n_l \rightarrow \lambda = -N$$ ← let us add up the above equation over all l and use the constraint

$$\rightarrow \pi_l^{ML} = \frac{n_l}{N}$$ ← eliminating λ we obtain the π_l of maximum likelihood!

ML for the Gaussian

- Another "easy" case for ML density estimation: the Gaussian pdf. Its parameters are θ = {μ, Σ}. The log-likelihood function for a D-dimensional variable is:

$$\ln\left(\prod_{i=1}^{N} N(x_i \mid \mu,\Sigma)\right) = \sum_{i=1}^{N} \ln N(x_i \mid \mu,\Sigma) = \sum_{i=1}^{N} \ln\left(\frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x_i-\mu)^T \Sigma^{-1}(x_i-\mu)}\right) =$$

$$= \sum_{i=1}^{N}\left(-\ln\left((2\pi)^{D/2}|\Sigma|^{1/2}\right) - \frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right) =$$

$$= -\frac{N}{2}\ln|\Sigma| - \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$$

- We must then maximise this function in μ, Σ jointly. In addition, we must guarantee that Σ be positive (semi)definite

ML for the Gaussian

- The likelihood function for the Gaussian allows for a global maximum. Moreover, the solution is in closed form:

$$\mu_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\Sigma_{ML} = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_{ML})(x_i-\mu_{ML})^T$$

The maximum-likelihood variance, Σ_{ML}, is a so-called *biased* estimate of the true variance; it should be multiplied by N/(N-1) to unbias; yet, the correction is negligible for any reasonable N

ML for mixture distributions

- Finding maxima of L(θ) for mixture distributions is not as easy as for the Gaussian. A popular approach – known as Expectation-Maximisation – will be presented in the following
- In addition, the likelihood function for mixture distributions generally has *multiple, local maxima*; we are not ensured that we can find the global maximum, or even a "good" one. A simple case for a Gaussian mixture model (GMM) is shown in the next few slides
- Certain maxima, even with infinite likelihood are just non desirable (overfitting)

The likelihood function for a GMM: an example

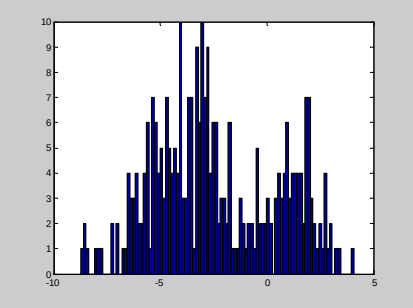
- Just an example to display the likelihood function for a simple GMM
- As we can easily visualise functions of 2 parameters, we choose the following simple 1-D model:

$$p(x) = 0.3\,N(x \mid \mu_1, \sigma_1 = 1.6) + 0.7\,N(x \mid \mu_2, \sigma_2 = 1)$$

  where the only parameters are μ_1 and μ_2
- μ_1 and μ_2 are made vary in range -10 ÷ +8 in 0.5 steps

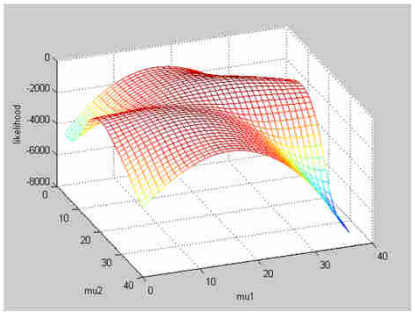
The data

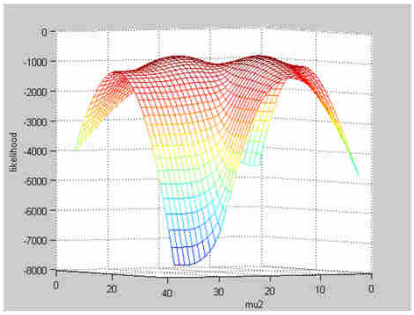
- 300 uni-dimensional samples
- Their histogram:



Likelihood surface

- Two maxima found, at:
  - $\mu_1 = 0.5$, $\mu_2 = -4.5$ (log-likelihood: -775.5742)
  - $\mu_1 = -4$, $\mu_2 = 1$ (log-likelihood: -807.7207)

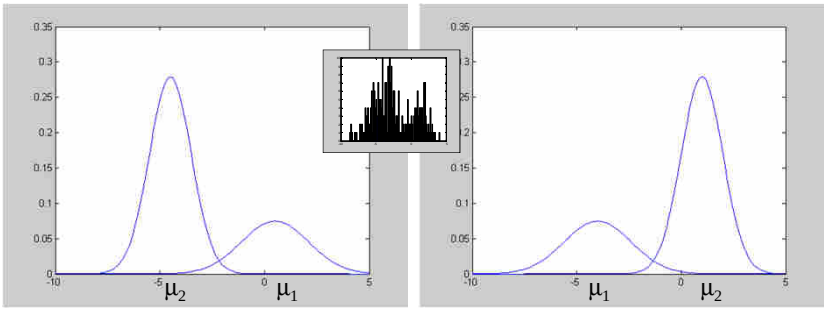




Quality of fitting

- One maximum is clearly better than the other

maximum at $\mu_1 = 0.5$, $\mu_2 = -4.5$ maximum at $\mu_1 = -4$, $\mu_2 = 1$



The EM approach

- Direct maximisation of the log-likelihood is often inconvenient or just difficult
- A very popular alternative for ML estimation is given by the Expectation-Maximization (EM) approach
- In the EM approach, instead of maximising function $L(\theta)$ (or $LL(\theta)$), we maximise another function, **$Q(\theta)$**
- The approach works since, for most models, obtaining a maximum for $Q(\theta)$ guarantees a maximum for in $LL(\theta)$ over an initially arbitrary choice of θ

- The approach was proposed by Dempster, Laird, Rubin (DLR) in "Maximum Likelihood from Incomplete Data via the EM Algorithm," Journal of the Royal Statistical Society, Series B, 1977 – a 22-page paper with 23 reviewers
- You can see H. Tagare, "A Gentle Introduction to the EM algorithm. Part I: Theory", http://noodle.med.yale.edu/hdtag/pubs/em.ps for an easy introduction

The EM approach

- EM posits the existence of latent variables, Y, and sets a different target for maximisation:

$$Q\left(\theta,\theta^{old}\right)=E\left[\ln p\left(X,Y\mid\theta\right)\mid X,\theta^{old}\right]=$$
$$=\int \ln p\left(X,Y\mid\theta\right)p\left(Y\mid X,\theta^{old}\right)dY$$

- L(θ) is therefore called the *incomplete data* likelihood, while $Q(\theta,\theta^{old})$ is the expected value of the *complete data* log-likelihood, ln p(X,Y|θ), on conditional probability $p(Y|X,\theta^{old})$
- For the approach to make sense, $Q(\theta,\theta^{old})$ must be such that its maximisation is easier than that of L(θ)

The EM approach: $Q(\theta,\theta^{old})$

Many important things can be said about EM; here we recap the main:

- For the approach to make sense, the expression for p(X,Y|θ) must be significantly simpler than that of p(X|θ); otherwise, we'd better maximise LL(θ) directly
- Finding an expression for $Q(\theta,\theta^{old})$ requires expressions for ln p(X,Y|θ) and $p(Y|X,\theta^{old})$, and the ability to integrate their product over Y
- If analytic integration cannot be used, Monte Carlo methods can be used instead (samples from $p(Y|X,\theta^{old})$) or variational approximations of $p(Y|X,\theta^{old})$ which factorise and integrate more nicely
- $Q(\theta,\theta^{old})$ must then be differentiated in θ and maxima found

The EM approach

- After the maximum of $Q(\theta,\theta^{old})$, θ^{new}, is found, the process starts a new iteration from $Q(\theta,\theta^{new})$. Iterations continue until convergence
- Each iteration **is guaranteed to increase (or not decrease) the incomplete-data log-likelihood, LL(θ|X)** - that is what we want
- The maximum we find in the parameter space upon convergence is a local one! Its position depends on the choice of the initial θ^{old}
- Each iteration consists of two steps:
  - the **E step**, where we compute the updated $p(Y|X, \theta^{old})$
  - the **M step**, where better θ are chosen by differentiation of $Q(\theta,\theta^{old})$
- A proof of EM is given in the Appendix

A more general framing: Neal, Radford; Hinton, Geoffrey (1999), "A view of the EM algorithm that justifies incremental, sparse, and other variants," Learning in Graphical Models, Michael I. Jordan. ed., Cambridge, MA: MIT Press, pp. 355–368

The EM algorithm

1. Choose an initial θ^{old}
2. E step: compute $p(Y|X,\theta^{old})$
3. M step: compute $Q(\theta,\theta^{old})$ and find its maxima, θ^{new}
4. Check for convergence of either L or θ; if not,

   $\theta^{old} \leftarrow \theta^{new}$

   and return to step 2

EM for GMM

- EM is the main tool to find maximum-likelihood parameters for a mixture distribution, including GMM
- EM for GMMs assumes that, for each x_i sample, there exists a latent discrete r.v., y_i, whose value, l={1..M}, is the index of the Gaussian component responsible for generating that sample
- It is assumed that each x_i depends only on its y_i and vice versa
- Therefore:

$$p(x_i, y_i) = p(x_i \mid y_i)\, p(y_i) = N(x_i \mid \mu_{y_i}, \Sigma_{y_i})\, \alpha_{y_i}$$

- Note that $p(x_i, y_i)$ is much simpler than $p(x_i)$, as we wanted:

$$p(x_i) = \sum_{y_i=1}^{M} \alpha_{y_i}\, N(x_i \mid \mu_{y_i}, \Sigma_{y_i})$$

EM for GMM

- The expression for ln p(X,Y| θ) is therefore:

$$ln\, p(X, Y \mid \theta) = ln \prod_{i=1}^{N} p(y_i, x_i \mid \theta) = \sum_{i=1}^{N} ln\, p(y_i, x_i \mid \theta) =$$

$$= \sum_{i=1}^{N} ln(\alpha_{y_i}\, N(x_i \mid \mu_{y_i}, \Sigma_{y_i}))$$

- The expression for $p(Y|X, \theta^{old})$ is:

$$p(Y \mid X, \theta^{old}) = \prod_{i=1}^{N} p(y_i \mid x_i, \theta^{old})$$

where $p(y_i \mid x_i, \theta^{old}) = \dfrac{\alpha_{y_i}^{old}\, N(x_i \mid \mu_{y_i}^{old}, \Sigma_{y_i}^{old})}{\sum_{y_i} \alpha_{y_i}^{old}\, N(x_i \mid \mu_{y_i}^{old}, \Sigma_{y_i}^{old})}$

- $Q(\theta, \theta^{old})$ multiplies these two terms and integrates over Y

EM for GMM

- With some manipulation (see, e.g., [Bilmes 98]), $Q(\theta,\theta^{old})$ becomes:

$$Q(\theta,\theta^{old}) = \sum_{l=1}^{M}\sum_{i=1}^{N} ln(\alpha_l\ N(x_i \mid \mu_l,\Sigma_l))p(y_i = l \mid x_i,\theta^{old}) =$$
$$= \sum_{l=1}^{M}\sum_{i=1}^{N} ln(\alpha_l)p(y_i = l \mid x_i,\theta^{old}) + \sum_{l=1}^{M}\sum_{i=1}^{N} ln(N(x_i \mid \mu_l,\Sigma_l))p(y_i = l \mid x_i,\theta^{old})$$

- $Q(\theta,\theta^{old})$ is then differentiated to find its maximum in θ; a constraint, $\Sigma_{l=1..M}\ \alpha_l = 1$, needs to be added to find meaningful weights, α_l
- Please note that the constrained maximum in α_l and those in μ_l, Σ_l are unique

EM for GMM: re-estimation formulas

- E step:

$$p(y_i = l \mid x_i,\theta^{old}) = \frac{\alpha_l^{old} N(x_i \mid \mu_l^{old},\Sigma_l^{old})}{\sum_{k=1}^{M}\alpha_k^{old} N(x_i \mid \mu_k^{old},\Sigma_k^{old})}$$

aka *responsibility* of component *l* for sample x_i

- M step:

$$\alpha_l^{new} = \frac{1}{N}\sum_{i=1}^{N} p(l \mid x_i,\theta^{old})$$

$$\mu_l^{new} = \frac{\sum_{i=1}^{N} x_i p(l \mid x_i,\theta^{old})}{\sum_{i=1}^{N} p(l \mid x_i,\theta^{old})}$$

$$\Sigma_l^{new} = \frac{\sum_{i=1}^{N}(x_i - \mu_l^{new})(x_i - \mu_l^{new})^T p(l \mid x_i,\theta^{old})}{\sum_{i=1}^{N} p(l \mid x_i,\theta^{old})}$$

GMM: size of parameters

- With D-dimensional data, the GMM parameters' size are as:
  - for each weight, α_l (aka $P(\omega_l)$ or π_l): a scalar
  - for each mean, μ_l: a D x 1 vector
  - for each covariance matrix, Σ_l: a D x D symmetric matrix
    - with D(D+1)/2 dof, if full
    - D, if diagonal
    - 1, if spherical
- At times, the covariances are constrained to be the same for all components
- Re-estimation formulas vary accordingly

EM for mixture models: caveats

- *Singularities* may arise during training:

  one component models one datum only, tightly

  $\rightarrow$ its covariance tends to 0
  $\rightarrow$ $p(x|\theta_l)$ tends to ∞
  $\rightarrow$ $p(x|\theta)$ also tends to ∞
  $\rightarrow$ $\ln p(x|\theta)$ also tends to ∞
  $\rightarrow$ $LL(\theta) = \Sigma_i \ln p(x_i|\theta)$ also tends to ∞

  we have reached a maximum of the likelihood; yet, the model's parametrisation is not useful

  Common trick: add some epsilon to the principal diagonal of Σ
- The local maximum we reach upon convergence may vary heavily with the initial parameters

ML and MAP density estimation

- **ML** density estimation:

$$\theta_{ML} = \arg\max_{\theta}\left(p\left(X \mid \theta\right)\right)$$

- **MAP** density estimation: think of parameters θ as r.v. themselves, allowing some prior distribution p(θ):

$$\theta_{MAP} = \arg\max_{\theta}\left(p\left(\theta \mid X\right) \propto p\left(X \mid \theta\right)p\left(\theta\right)\right)$$

  - MAPE is useful to favour certain values of θ
  - Intending to apply EM, one can see that an ln p(θ) term must be added to Q(θ,θ*). The M step must now maximise: Q(θ,θ*) + ln p(θ). Argmax (Q(θ,θ*) + ln p(θ)) is of course generally different from argmax (Q(θ,θ*)) and argmax(ln p(θ)), and may not be easy. Please note that function ln p(θ) does not vary during iterations.

Posterior maximisation by sampling

- Like for the likelihood p(X| θ), it is also possible to maximise p(θ|X) in θ directly, without passing through Bayes' theorem and EM
- *Posterior sampling* techniques sample p(θ|X) many times and compute the maximum from the histogram; sampling is possible because p(θ|X) is a distribution in θ
- In alternative, one can obtain *sequential* (non i.i.d) *samples* from p(θ|X) with Markov chain Monte Carlo (MCMC) techniques which are much more efficient than i.i.d. sampling

Penalised maximum likelihood

- To avoid overfitting, the likelihood function can be traded off with some “penalty factor” (aka regularizer)
- MAP estimation can also be seen as a form of penalised maximum likelihood where the penalty is incurred when moving away from the prior
- Examples:

$$\arg\max_{\theta}\left(\sum_{i=1}^{N} log\, p(x_i \mid \theta) - \lambda\|\theta\|_2^2\right) \quad (\lambda > 0)$$ encourages *small* values for θ

$$\arg\max_{\theta}\left(\sum_{i=1}^{N} \log p(x_i \mid \theta) - \lambda\|\theta\|_1\right) \quad (\lambda > 0)$$ encourages *sparse* values for θ

Bayesian predictive distribution

- Both ML and MAP are “point estimates” of the parameters used in the predictive distribution. It is also possible to deal with θ more fully by marginalising it as in:

$$p(x \mid X) \approx \int p(x \mid \theta) p(\theta \mid X)\, d\theta$$

- At times, the above integration can be done in closed-form and is easily manageable
- “Such marginalizations lie at the heart of Bayesian methods for pattern recognition” (C. Bishop)

Non-parametric estimators

- GMMs belong to the general category of parametric density estimators
- A different approach to density estimation can be taken by choosing models with a minimal number of parameters
- Widespread approaches include:
  - histograms
  - *k*-nearest neighbours (kNN)
  - kernel density estimation (KDE)
  - mean-shift vector

Histogram

- With the histogram, the data space is divided in a regular grid (each element is called a bin)
- p(x) is uniform within each bin and given by:

  (*number of samples in the bin*)/(*total number of samples*)
- Limitations:
  - sharp/non-smooth estimate
  - depends on the size of the bins
  - depends on the alignment of the grid
  - number of bins grows exponentially with D
- Useful for visualization in 1 or 2D

Generic non-parametric estimation

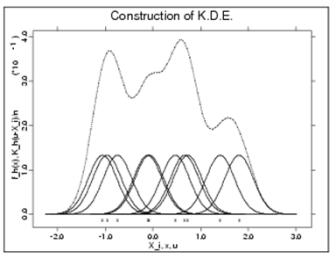
- p(x) ~ k/NV, where
  - V is the volume surrounding x
  - k is the number of samples in V
  - N is the total number of samples

  it provides a good estimate if N is large, k grows with N and V is small enough for p(x) to be constant
- Two main approaches, KDE and kNN
  - in KDE, V is fixed and k computed from data set
  - in kNN, k is fixed and V computed from data set
- The *mean shift vector* approach is a post processing of KDE that finds the distribution's modes explicitly; ends up in a mixture model, but in a non-parametric manner

Kernel Density Estimation

- A kernel function, K(u) (aka Parzen window), is fit centred on each sample
- Typical kernels (in 1D):

Uniform	$K(u) = 1/2 \quad \lvert u \rvert \le 1$
Triangle	$K(u) = (1 - \lvert u \rvert) \quad \lvert u \rvert \le 1$
Epanechnikov	$K(u) = \frac{3}{4}(1 - u^2) \quad \lvert u \rvert \le 1$
Gaussian	$K(u) = (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}u^2\right)$



KDE pdf

- Kernels are then all added up and sum normalised; this is the KDE pdf (in D dimensions):

$$p(x) = \frac{1}{Nh^D} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h} \right)$$

  h is called the *bandwidth*

- Kernels are typically radially symmetric, so there is only one scalar parameter, h, also in D dimensions; it equates to a spherical covariance matrix
- Given x, evaluation of p(x) is computationally heavy: high execution time

KDE bandwidth

- How to choose the bandwidth?
  - Maximum likelihood would lead to a useless solution:
    h = 0
  - A *pseudo*-likelihood can be used in place of the standard likelihood:
    when evaluating $p(x_i)$ in $L(\theta)$, leave the kernel centred on it out; in this way, sample x_i has to be "explained" by its closest neighbours
- Many other methods to estimate the bandwidth:
  - Maximal Smoothing Principle, Least Squares Cross Validation, Biased Cross Validation, Smoothed Cross Validation, … many!

    B. A. Turlach. Bandwidth Selection in Kernel Density Estimation: A Review. Technical Report Université Catholique de Louvain, Belgium, 1993

GMM vs KDE

- GMM

$$p_{GMM}(x) = \sum_{l=1}^{M} \alpha_l N(x \mid \mu_l, \Sigma_l)$$

- KDE (Gaussian kernel)

$$p_{KDE}(x) = \frac{1}{N} \sum_{i=1}^{N} N(x \mid x_i, \Sigma)$$

one component per observation!

only one Σ for all components, and spherical

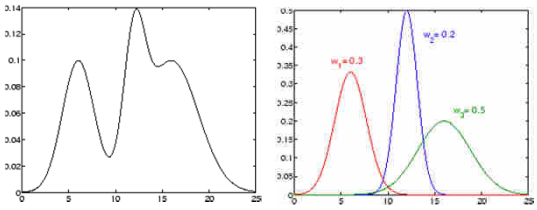
equal weights

centred in the observation

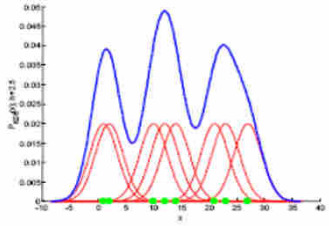
- Similarity only notational

GMM vs KDE

- GMM



- KDE (Gaussian kernel)

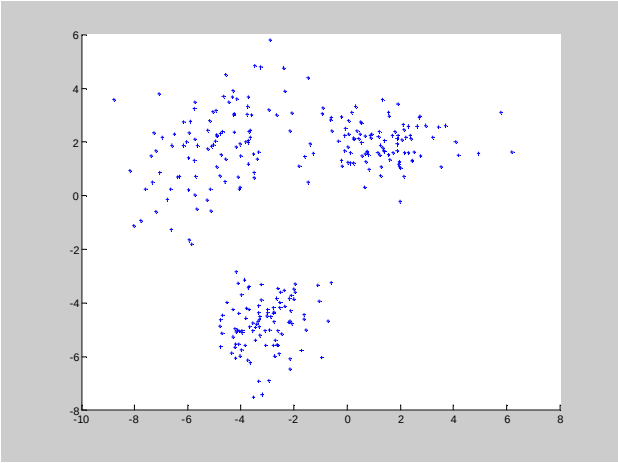


GMM vs KDE: example

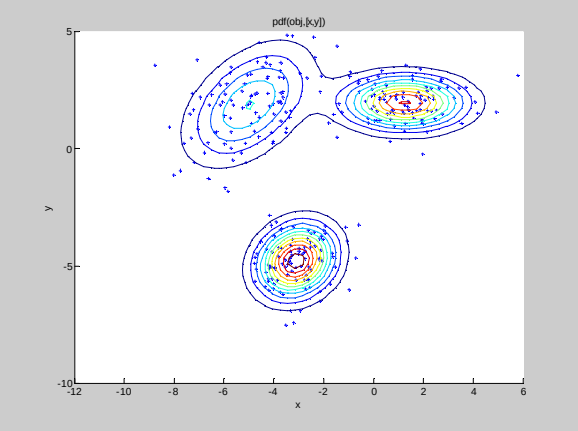
- Example with 3 modes in 2D
- Parameters in GMM with different constraints on covariance; how model changes
  - full: 2 + 3 * 2 + 3 * 3 = 17 parameters
  - diagonal: 2 + 3 * 2 + 3 * 2 = 14 parameters
  - spherical: 2 + 3 * 2 + 3 * 1 = 11 parameters
  - shared full: 2 + 3 * 2 + 1 * 3 = 11 parameters
  - shared diagonal: 2 + 3 * 2 + 1 * 2 = 10 parameters
  - shared spherical: 2 + 3 * 2 + 1 * 1 = 9 parameters
- Spherical model not as restrictive for KDE
  - spherical kernel: 1 parameter

Speaker's notes

The data

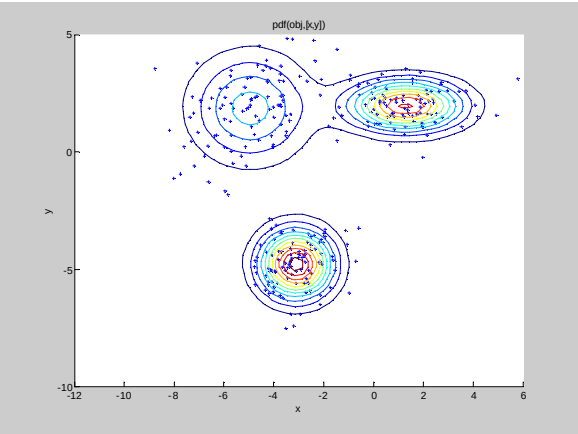


GMM, full covariance



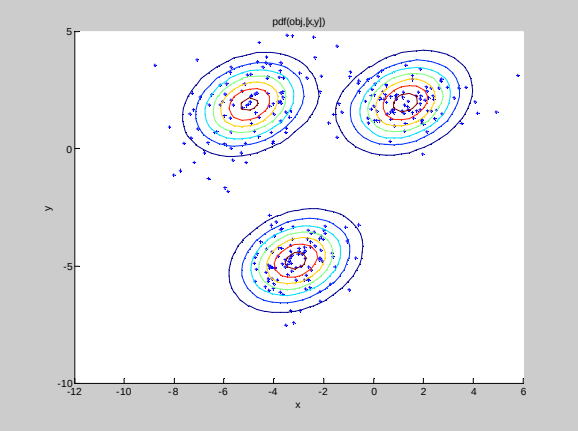
- log-likelihood = -1241.64

GMM, diagonal covariance



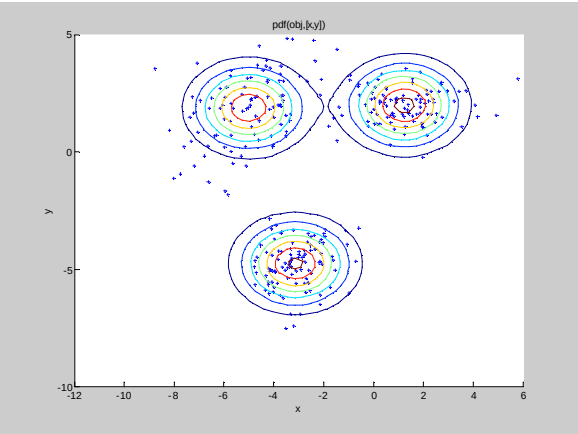
- log-likelihood = -1253.88

GMM, shared full covariance



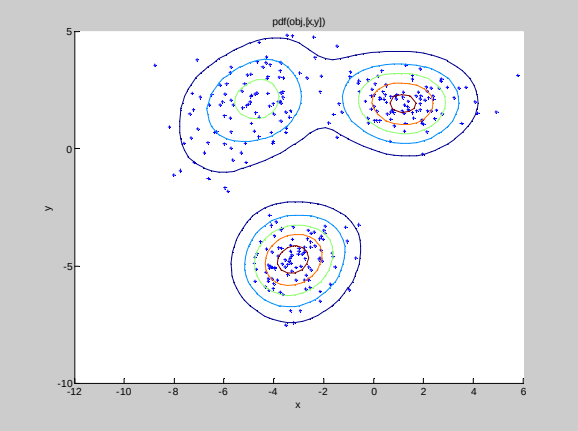
- log-likelihood = -1283.81

GMM, shared diagonal covariance



- log-likelihood = -1291.52

KDE, spherical Gaussian kernel



- only 1 free parameter! But all data will be needed for predictions
- log-likelihood compares with GMM

Example papers

- Application: *background subtraction*

  Extracting moving objects in a video

- Mixture models:

  C. Stauffer and W.E.L. Grimson, “Adaptive background mixture models for real-time tracking,” Proc. IEEE CVPR 1999, pp. 246–252.

  1317 cites on Google Scholar (19 Nov 08). Many modification papers have followed.

- KDE:

  A. Elgammal, D. Harwood, and L.S. Davis, “Non-parametric model for background subtraction,” Proc. ECCV 2000, pp. 751-767.

Appendix: the logical implant of EM

- EM consists of various steps and can be presented in many different ways; here, we follow the simple, yet elegant presentation of [Tagare 1998]. For greater detail and comprehensiveness, we recommend [Neal & Hinton 1998]

1. The first step is to write the log-likelihood, LL(θ), as the sum of two functions:

$$LL(\theta) = Q(\theta;\theta^0) + H(\theta;\theta^0)$$

parametric in an **arbitrary parameter**, say, θ^0. The value of these functions will be disclosed later; at this stage, all that is important is that parameters θ and θ^0 are homogenous, in the sense that they are composed of the same set of parameters

The logical implant of EM (2)

2. The next step is to note that function $H(\theta; \theta^0)$ has a minimum for $\theta = \theta^0$; to be shown later
3. We can now write the LL(θ) function in the specific argument θ^0, same as the arbitrary parameter:

$$LL(\theta^0) = Q(\theta^0;\theta^0) + H(\theta^0;\theta^0)$$

4. The question we ask is: given log-likelihood value $LL(\theta^0)$, is it generally possible to find $\theta^1 : LL(\theta^1) \geq LL(\theta^0)$?
5. By using the given decomposition, the above translates into:

$$Q(\theta^1;\theta^0) + H(\theta^1;\theta^0) \geq Q(\theta^0;\theta^0) + H(\theta^0;\theta^0)$$

The logical implant of EM (3)

6. or:

$$Q\left(\theta^1;\theta^0\right) \geq Q\left(\theta^0;\theta^0\right) - \overbrace{\left(H\left(\theta^1;\theta^0\right) - H\left(\theta^0;\theta^0\right)\right)}^{\geq 0}$$

7. Given that $H(\theta; \theta^0)$ has a minimum for $\theta = \theta^0$, a sufficient condition for the above is:

$$Q\left(\theta^1;\theta^0\right) \geq Q\left(\theta^0;\theta^0\right)$$

which can be guaranteed, for instance, by setting θ^1 to:

$$\theta^1 = \arg\max_{\theta} Q\left(\theta;\theta^0\right)$$

or any other value satisfying the last inequality (this is called "generalised" EM, or GEM). Often, the above maximisation is easy and the function unimodal; it is, in general, a maximisation

The logical implant of EM (4)

8. We have therefore proven that satisfying condition $Q(\theta^1; \theta^0) \geq Q(\theta^0; \theta^0)$ guarantees that $LL(\theta^1) \geq LL(\theta^0)$. This basically leads us to the end of our description. We can now write $LL(\theta^1)$ around θ^1 as the arbitrary parameter:

$$LL\left(\theta^1\right) = Q\left(\theta^1;\theta^1\right) + H\left(\theta^1;\theta^1\right)$$

and find $\theta^2 : LL(\theta^2) \geq LL(\theta^1)$ in a similar way.

9. The process is repeated by induction and should converge at a local maximum of $LL(\theta)$

LL = Q + H

$$LL(\theta) =: \ln p(X \mid \theta) = \ln\left(\frac{p(X,Y \mid \theta)}{p(Y \mid X,\theta)}\right) = \ln p(X,Y \mid \theta) - \ln p(Y \mid X,\theta)$$

- We can now integrate over a density on Y, $p(Y|X,\theta^0)$:

$$\int_Y \ln p(X \mid \theta)\, p(Y \mid X,\theta^0)\, dY = \ln p(X \mid \theta) =$$

$$\int_Y \ln p(X,Y \mid \theta) p(Y \mid X,\theta^0)\, dY - \int_Y \ln p(Y \mid X,\theta) p(Y \mid X,\theta^0)\, dY$$

because the first member, $p(X|\theta^0)$ does not depend on Y

LL = Q + H (2)

- Q and H are thus defined as follows:

$$Q(\theta;\theta^0) \equiv \int_Y \ln p(X,Y \mid \theta) p(Y \mid X,\theta^0)\, dY$$

$$H(\theta;\theta^0) \equiv -\int_Y \ln p(Y \mid X,\theta) p(Y \mid X,\theta^0)\, dY$$

- Neal & Hinton show that choosing density $p(Y|X,\theta^0)$ in the space of densities on Y, q(Y), is a maximisation at its turn (E step of EM)

$$H(\theta^1;\theta^1) \geq H(\theta^0;\theta^0)$$

$$H(\theta^1;\theta^0) - H(\theta^0;\theta^0) =$$

$$= \int_Y \left[-\ln p(Y \mid X, \theta^1) + \ln p(Y \mid X, \theta^0) \right] p(Y \mid X, \theta^0) dY =$$

$$= \int_Y \left[-\ln \frac{p(Y \mid X, \theta^1)}{p(Y \mid X, \theta^0)} \right] p(Y \mid X, \theta^0) dY \geq$$

$$(Jensen's\ inequality) \geq -\ln \int_Y \left[\frac{p(Y \mid X, \theta^1)}{p(Y \mid X, \theta^0)} \right] p(Y \mid X, \theta^0) dY =$$

$$= -\ln \int_Y p(Y \mid X, \theta^1) dY = -\ln 1 = 0$$

