# Bayesian Machine Learning - Lecture 5

Guido Sanguinetti

Institute for Adaptive and Neural Computation
School of Informatics
University of Edinburgh
gsanguin@inf.ed.ac.uk

March 2, 2015

# Today's lecture

# Supervised, unsupervised and semi-supervised learning

- Discriminative vs generative learning partitions learning techniques in terms of how predictions are formulated
- One can also partition learning methods in terms of what kind of information is available
- When the data consists of pairs $(\mathbf{x}, y)$, we talked of *supervised learning* (learning a map)
- When the data consists only of variables of the same type $\mathbf{x}$ and we are interested in recovering structure from the data, we talk of *unsupervised learning*
- In a hybrid scenario where we have some pairs and some inputs on their own, we talk of *semi-supervised learning*
- Often input data are cheap and plentiful, output not so

# Unsupervised learning

- Unsupervised learning consists in finding structure, or patterns, in data
- Discuss possible examples

# Latent variables

- E1: data is human height, population structure gender
- E2: data is gene expression time series ($\sim$20K dimensional), structure given by physiological processes going on ($\sim$50)
- In both cases, the observed data distribution is the *marginal* of a joint distribution of variables characterising the population structure (gender, physiology) and observed
- These are examples of **LATENT VARIABLE MODELS**

# Main example: Gaussian mixture models

- Classic model used in clustering and density modelling
- The marginal density is given by

$$p(\mathbf{x}|\Theta) = \sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\mu_j, \Sigma_j) \tag{1}$$

  where $\Theta$ denotes collectively the parameters $\pi_j$ (*mixing coefficients*) and $\mu_j, \Sigma_j$
- The latent variable interpretation is obtained by introducing categorical variables $z \in \{0,1\}^K$ (membership variables) and identify $\pi_j = p(z_j = 1)$
- Learning can be done by ML, either directly through gradient methods or using the Expectation-Maximisation algorithm (rest of today)

# Assessing unsupervised learning results

- Tricky, as there is no ground truth (as in supervised learning)
- One can keep some data aside and evaluate the likelihood of the new data under the learnt model (ideally should be high)
- Plausible but in my opinion very weak
- One can compare different models using information criteria such as the *Akaike Information Criterion* (AIC)

$$AIC = 2k - 2\log(\mathcal{L})$$

and the *Bayesian Information Criterion* (BIC)

$$BIC = k\log(n) - 2\log(\mathcal{L})$$

- Evaluate variance reduction, i.e. compare $var[\mathbf{x}]$ with $E_{p(z|\mathbf{x})}[var[\mathbf{x}|z]]$

# Jensen's inequality

- The EM algorithm relies on *Jensen's inequality*, a general result on the integrals of convex functions
- The form that we are concerned about is the following: let $\phi\colon \mathbb{R}^d \to \mathbb{R}$ be a concave function, and let $p$ be a probability distribution
- Then

$$E_p[\phi(\mathbf{x})] \leq \phi\left(E_p[\mathbf{x}]\right) \tag{2}$$

- Let's prove it with a picture when $p$ is a finite distribution

# EM in a nutshell

- The EM algorithm uses Jensen's inequality (2) to determine a (family of) lower bound on the log marginal likelihood

$$\log p(\mathbf{x}|\Theta) = \log \sum_z p(\mathbf{x}, z|\Theta) = \log \sum_z \frac{p(\mathbf{x}, z|\Theta)}{q(z)} q(z) \geq \sum_z q(z) \log \frac{p(\mathbf{x}, z|\Theta)}{q(z)} = \mathcal{L}_q(\mathbf{x}, \Theta) \tag{3}$$

  where $q(z)$ is *any* distribution over $z$

- The lower bound is saturated (i.e. $\mathcal{L}_q(\mathbf{x}, \Theta) = \log p(\mathbf{x}|\Theta)$) if and only if $q(z) = p(z|\mathbf{x}, \Theta)$

# EM in a nutshell

The EM algorithm then proceeds as follows

1. Initialise the parameters $\Theta$
2. For fixed $\Theta$, compute the posterior $p(z|\mathbf{x}, \Theta)$ and use it to compute the bound $\mathcal{L}_q(\mathbf{x}, \Theta)$ (E-step)
3. Maximise the bound w.r.t. $\Theta$ (M-step)
4. If not converged, return to step 2

Each EM step will increase the log-likelihood (why?)$\longrightarrow$ EM is guaranteed to converge to a local optimum

# EM for Gaussian Mixture Models: E-step

- The joint probability for $z = j$ and $\mathbf{x}$ according to (1) is

$$p(z = j, \mathbf{x}) = \pi_j \mathcal{N}(\mathbf{x}|\mu_j, \Sigma_j)$$

- By Bayes' theorem, the posterior is proportional to the joint with constant given by the marginal

$$\gamma_{ij} = \frac{\pi_j \mathcal{N}(\mathbf{x}_i|\mu_j, \Sigma_j)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_i|\mu_j, \Sigma_j)}$$

- $\gamma_{ij}$ are called *responsibilities*: they describe what is the probability that cluster $j$ is responsible for data point $i$

# EM for Gaussian Mixture Models: E-step

- Recall that a discrete probability can be written succinctly as

$$p(z) = \prod_j \pi_j^{z_j}$$

- Then the joint log likelihood for all $N$ points can be written as

$$\begin{aligned} \mathcal{L}(\mathbf{x}, z|\Theta) &= \log \left[ \prod_{i=1}^{N} \prod_{j=1}^{K} \left(\pi_j \mathcal{N}\left(\mathbf{x}_i|\mu_j, \Sigma_j\right)\right)^{z_{ij}} \right] = \\ &= \sum_{i=1}^{N} \sum_{j=1}^{K} \left[ z_{ij} \left( \log \pi_j + \log \mathcal{N}\left(\mathbf{x}_i|\mu_j, \Sigma_j\right)\right)\right] \end{aligned} \tag{4}$$

- To compute its expectation w.r.t. the posterior, just replace $z_{ij}$ with the responsibilities $\gamma_{ij}$

# EM for Gaussian Mixture Models: M-step

- The lower bound computed in (4) is very easy to optimize (sum of quadratics)
- The M-step equations can be solved *analytically* and have very pleasing interpretations, e.g.

$$\mu_j = \sum_{i=1}^{N} \gamma_{ij} \mathbf{x}_i$$

- Work them all out as an exercise, and remember $\sum \pi_j = 1$!

# EM pros and cons

- Nice, interpretable and tractable equations
- Usually fast, few iterations land you in the optimum
- Vulnerable to local optima
- Only returns point estimates of the parameters (no measure of associated uncertainty)
- For complex models, dependence on initialisation can make it virtually useless

# Supervised Gaussian mixture models

- Of course, you can also use a Gaussian mixture model if you know the class variables $z$
- In that case, learning can be done just via the M-step (replacing the responsibilities with the actual 0-1 labels)
- Example: consider the simpler case with only two classes with the same variance
- Exercise: show that the posterior probability of being in one class fulfils a logistic regression equation
- How many parameters do you need for the generative representation? How many for LR?

# Partial supervision

- One can also consider the hybrid situation when some labels are available and some (most) are not
- The log likelihood is given by a sum of log likelihoods, one from the supervised part, and one for the unsupervised
- Learning can be performed by a modified EM algorithm, where we apply the E-step only to the unsupervised part, and the M-step on the whole likelihood
- Potentially problematic when one has a lot more unlabelled data than labelled data, as the label information can get swamped
- Reweighting methods exist where the supervised log-likelihood is rescaled to increase its importance

# Mixtures of discriminative models

- So far we have considered the purely generative scenario where we had data only of "input" type
- One could equally well consider a scenario where input-output pairs come from a mixture
- The generative scenario is that first a particular component is chosen, then an input-output pair is assigned to a component
- Traditionally, discriminative models are called *experts*, hence the name mixture of experts

# Mixture of linear regression models

- Simplest mixture of experts model
- Conditioned on a discrete random variable $z$, we have

$$p(y|x,z) = \mathcal{N}(A_z x + b_z, \sigma_z^2)$$

- Introducting $\pi_i = p(z = i)$, we have a marginal

$$p(y|x) = \sum_{i=1}^{K} \mathcal{N}(A_z x + b_z, \sigma_z^2)$$

- Hence, the data-generating distribution consists of Gaussian distributed points about a number of different lines
- Exercise: work out the EM algorithm for mixture of linear regression experts