ST451 - Lent term
Bayesian Machine Learning

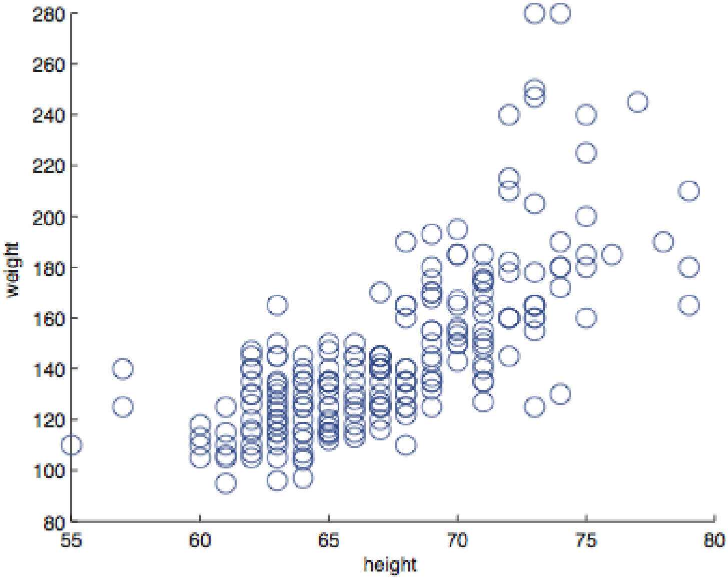
Kostas Kalogeropoulos

Clusters/Mixture Models and EM algorithm

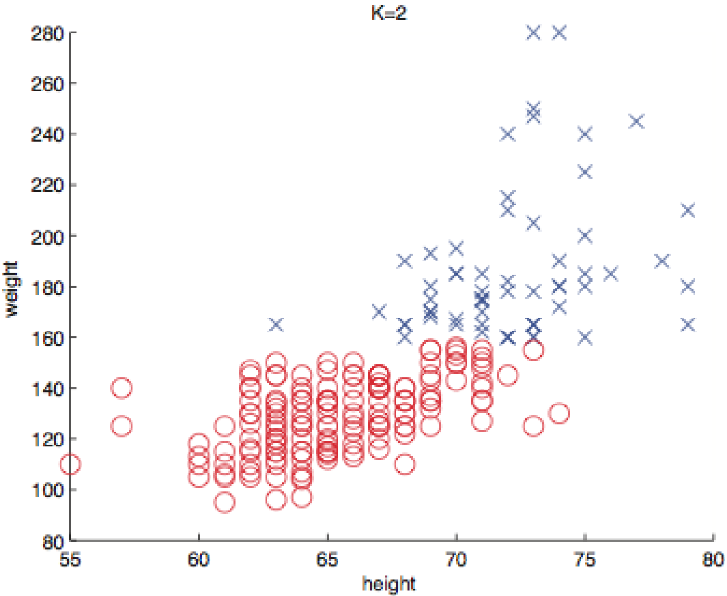
Outline

Outline

Motivating Example 1: Heights and weights

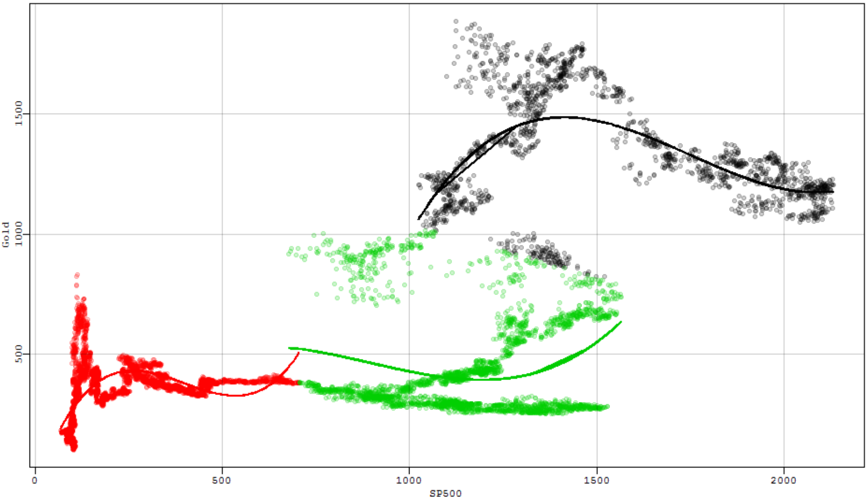


Example 1: Heights and weights



Example 2: Job Quality Definition/Measurement

Example 3: Financial Modelling

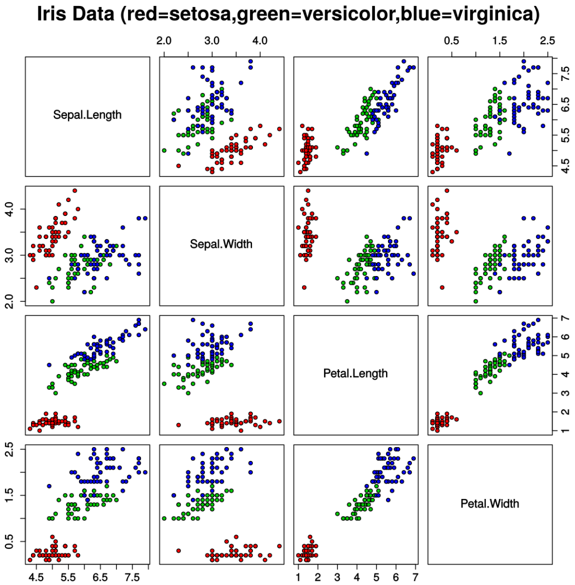


Example 4: Image Compression

Example 5: Distinguishing Iris flower species



Example 5: Distinguishing Iris flower species



Example 6: Bayesian non-parametric density models

Recall the VIX index and the model we used to capture some of its stylised facts

$$Y_t = Y_{t-1} + \kappa(\mu - Y_{t-1})\delta + \sigma\epsilon_t,$$

where Y_t is VIX at time t, δ is the time interval, and ϵ_t are independent error terms.

The distribution of each ϵ_t may assumed to be a mixture of Normal distributions.

Such model is very flexible; in this case corresponds to a jump diffusion model.

Outline

Data augmentation

Often we want to draw inference on parameters θ based on data x from a likelihood $f(x|\theta)$ that is either intractable or expensive to compute.

Introduce an unobserved latent variable z to extend the model defining $f(z, x|\theta)$

We can then work directly with $f(z, x|\theta)$ (variational Bayes, MCMC) or approximate the integral $f(x|\theta) = \int f(z, x|\theta)dz$ in some way (simulated likelihood, EM).

Many famous examples, e.g. Ising model, factor analysis, random effects, hidden Markov models and mixtures.

Cluster/mixture analysis

- The populations consists of K clusters/groups, each with distribution $f(x_i|\theta_k)$, $k = 1, \ldots, K$.
- Each individual $i = 1, \ldots, n$, belongs to one of these K clusters.
- Cluster indicator $z_i = 1, \ldots, K$ is an unobserved/latent categorical variable with Multinoulli distribution $\pi(z|\pi_k)$, where $\sum_k \pi_k = 1$.
- The aim is to classify individuals, by and draw inference on $\theta = (\pi_k, \theta_k)_{k=1}^K$.

Likelihood and augmented likelihood

Define also the z_{ik} indicator that takes the value 1 if the individual i is in cluster k and 0 otherwise. So if $z_i = 2$, $z_{i2} = 1$ and $z_{ik} = 0$ for $k \neq 2$.

The augmented likelihood also includes z_i for each x_i.

$$f(z_i, x_i|\theta) = \pi(z_i|\pi_k)f(x_i|z_i, \theta_k) = \prod_{k=1}^{K} \pi_k^{z_{ik}} f(x_i|\theta_k)^{z_{ik}}.$$

Note that $f(z_i = k, x_i|\theta) = \pi_k f(x_i|\theta_k)$. Summing out z_i gives

$$f(x_i|\theta) = \sum_{k=1}^{K} f(z_i = k, x_i|\theta) = \sum_{k=1}^{K} \pi_k f(x_i|\theta_k)$$

Overall we have $f(x|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$ and $f(z, x|\theta) = \prod_{i=1}^{n} f(x_i, z_i|\theta)$

Example: Gaussian Mixture Models

In Gaussian mixture models, we have $x_i | z = k \sim N(\mu_k, \Sigma_k)$

Hence

$$f(x_i|\theta) = \sum_{k=1}^{K} \pi_k N(x_i|\mu_k, \Sigma_k)$$

so the parameters to be estimated are $\theta = (\pi_k, \mu_k, \Sigma_k)_{k=1}^{K}$.

Due to the large number of parameters, especially for large K, restrictions are often placed on Σ_k, e.g. diagonal or tied.

Outline

Main idea

Complete Data: If we knew the cluster each person is, z_i, then MLE is straightforward: split the data into clusters do MLE in each cluster separately.

But we don't, so we need a modified approach. The algorithm used most frequently is the EM.

A rough sketch is the one below

1. Start with a θ.
2. E step: Use Bayes theorem to find the responsibilities $\gamma_{ik} = \pi(z_i = k|x, \theta)$ to get the expected log likelihood.
3. M step: Maximise the expected log-likelihood and update θ.
4. Continue until convergence.

log-likelihood and augmented log-likelihod

First write down the augmented log-likelihood. Remember that

$$f(z_i, x_i|\theta) = \prod_{k=1}^{K} \pi_k^{z_{ik}} f(x_i|\theta_k)^{z_{ik}},$$

so considering all individuals and taking log gives

$$\log f(z, x|\theta) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \left(\log \pi_k + \log f(x_i|\theta_k)\right)$$

By contrast the log-likelihood is

$$\log f(x|\theta) = \log \left[\sum_{k=1}^{K} \pi_k f(x_i|\theta_k)\right]$$

Notes

1. Can view the augmentation as way to bring log within the sum.
2. Easy to maximise the augmented log-likelihood given the z_{ik}'s.

E step

In the EM algorithm we update θ^{old} to θ^{new}. In the E step we define the expected log likelihood

$$Q(\theta, \theta^{old}) = \mathbb{E}_{\pi(z|x,\theta^{old})}\left[\sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\left(\log \pi_k + \log f(x_i|\theta_k)\right)\right]$$

Note that

$$\mathbb{E}_{\pi(z|x,\theta^{old})}[z_{ik}] = \frac{\pi_k^{old} f(x|\theta_k^{old})}{\sum_{j=1}^{K} \pi_j^{old} f(x|\theta_j^{old})} = \gamma(z_{ik}),$$

where the $\gamma(z_{ik}) = \pi(z_{ik} = 1|x, \theta^{old})$ are known as the responsibilities.

Hence we can write

$$Q(\theta, \theta^{old}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \gamma(z_{ik})\left(\log \pi_k + \log f(x_i|\theta_k)\right)$$

M step

The M step: consists of simply maximising $Q(\theta, \theta^{old})$ wrt to θ. Note that the $\gamma(z_{ik})$ are known numbers based on x and θ^{old} so it is usually an easy task.

To maximising $Q(\theta, \theta^{old})$ wrt to π_k's we can use Lagrange multipliers to satisfy the restriction that they sum to one. So we set

$$L = Q(\theta, \theta^{old}) + \lambda \left(\sum_k \pi_k - 1 \right),$$

$$\frac{\partial L}{\partial \pi_k} = 0 \quad \leftrightarrow \quad \pi_k = \frac{\sum_i \gamma(z_{ik})}{-\lambda}$$

$$\frac{\partial L}{\partial \lambda} = 0 \quad \leftrightarrow \quad \sum_k \pi_k = 1 \leftrightarrow \lambda = -n$$

so we get that $Q(\theta, \theta^{old})$ is maximised at

$$\pi_k^{new} = \frac{\sum_i \gamma(z_{ik})}{n} = \frac{n_k}{n}$$

Example: Gaussian Mixure models

The remaining parameters depend on which type of $f(x_i|\theta_k)$ we have.

For Gaussian mixture models standard MLE methods provide

$$
\begin{aligned}
\mu_k^{new} &= \frac{\sum_i \gamma(z_{ik})x_i}{\sum_i \gamma(z_{ik})} = \frac{\sum_i \gamma(z_{ik})x_i}{n_k} \\
\Sigma_k^{new} &= n\frac{1}{n_k}\sum_i \gamma(z_{ik})(x_i - \mu_k^{new})(x_i - \mu_k^{new})^T
\end{aligned}
$$

Hence the EM algorithm initiates θ and iteratively updates from θ^{old} to θ^{new} until the log likelihood or the parameters converge.

Similar results exist for other distributions such as Bernoulli, Exponential etc.

Connection with k-means

- Mixture models classify individuals to clusters based on the responsibilities $\gamma(\zeta_{ik})$'s, i.e. the posterior probabilities of z, rather than with certainty, aka soft allocation.

- This is reflected on the estimate of θ_k that are weighted averages based on how likely an individual is in cluster k.

- In Gaussian mixture models if we set $\Sigma_k = \sigma^2 I_d$ and let $\sigma^2 \to 0$ we get the same solution as with the k-means approach for μ_k. Note that in this case we have hard allocation.

- If we have general $\Sigma_k' s$ the approach coincide with the elliptical k-means.

Selecting the number of clusters

- In both mixture models and k-means it is not easy to select the number of classes.
- The default criterion in the mixture models is the BIC.
- Nevertheless the approach is very sensitive to starting values as the objective is multimodal and is very likely to get trapped in local maxima.
- It is recommended to initialise parameters based on intuition, try out multiple starting points or initialise with the results of another method.

Outline

Fully Bayesian approach

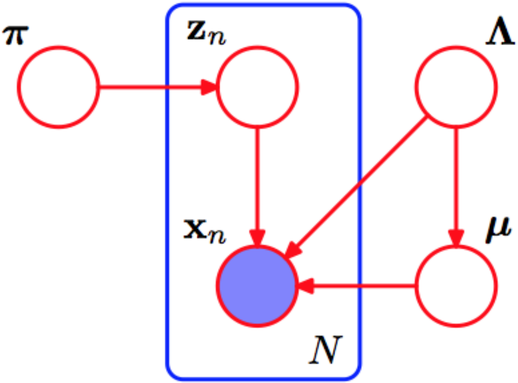
The approach so far was Bayesian with respect to z but not θ. For a fully Bayesian approach priors on θ should be specified.

In Gaussian mixture models example the conjugate priors can be used

$$
\begin{aligned}
\mu_k | \Lambda_k &\sim N\left(m_0,\ (\beta_0 \Lambda)^{-1}\right) \\
\Lambda &\sim \text{Wishart}\,(W_0, \nu_0) \\
\pi &\sim \text{Dirichlet}(\alpha_0)
\end{aligned}
$$

The posterior is not available in closed form. We can therefore consider a variational approximation.

Graph of Bayesian Gaussian Mixtures Model



Variational Inference for Gaussian Mixtures

We can apply mean field approximation of the form

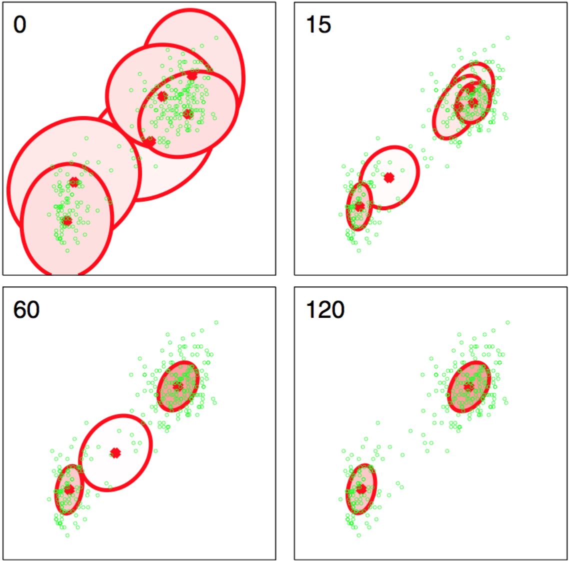
$$q(z, \pi, \mu, \Lambda) = q(z)q(\pi, \mu, \Lambda)$$

The outcome is analogous to the EM case.

1. $q(z)$ requires some tedious calculations that provide the responsibilities r_{ik}'s
2. Given r_{ik}'s $q(\pi, \mu, \Lambda)$ can be recognised as a product of a Dirichlet, Normal and Wishart distributions, the moments of which can be used to update r_{ik}'s. See Bishop 478-79 for details.
3. The above are repeated until convergence.

The prior parameter α_0, in the Dirichlet(α_0) prior on π, is critical in selecting the number of clusters. Large values favour equal π_k's but under values close to 0 only the important clusters will get individuals.

Variational Inference - no overfit



Latent Dirichlet allocation

Consider M documents each with N_i words. The model is defined as:

- Each of word w_{ij}, $i = 1, \ldots, M$ and $j = 1, \ldots, N_i$ may belong to different topics $k = 1, \ldots, K$, that indicate the word distribution ϕ_k:

$$w_{ij}|z_{ij} \ \sim \ \text{Multinoulli}(\phi_k)$$

- Let z_{ij} denote the topic indicator (latent) of w_{ij}, with distribution θ_i varying across documents:

$$z_{ij} \sim \text{Multinoulli}(\theta_i)$$

- The word and topic distributions (θ_i, ϕ_k) are given Dirichlet distribution priors with α, β respectively, both < 1 for sparsity:

$$\theta_i \sim \text{Dirichlet}(\alpha) \qquad \phi_k \sim \text{Dirichlet}(\beta)$$

Today's lecture - Reading

Bishop: 9.1 to 9.4, 10.2.1 10.2.2

Murphy: 11.1 11.2 11.4.1 11.4.2 21.6 27.3