Lecture 13:

Bayesian Parameter Estimation

Sam Roweis

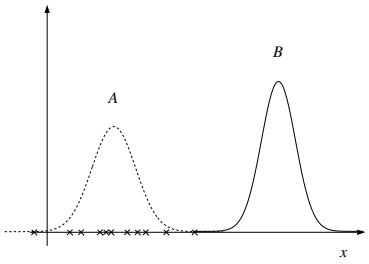
February 24, 2003

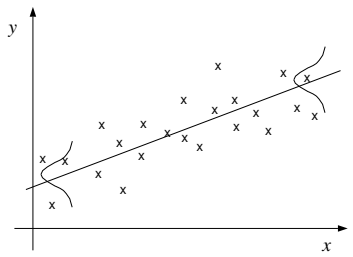
Bayesian Approach

- The Bayesian programme (after Rev. Thomas Bayes) treats *all unknown quantities as random variables* and represents uncertainty over those quantities using probability distributions.
- Thus, unknown parameters are treated as random variables just like latent (hidden) variables or missing data.
  This means we have probability distributions over the parameters.
- We can (and should) put priors $p(\theta)$ over them, and can compute things like posteriors $p(\theta|\mathcal{D})$.
- Sometimes we even want to integrate/sum them out as we did with things like cluster assignment variables or continuous factors.

Probability vs. Statistics

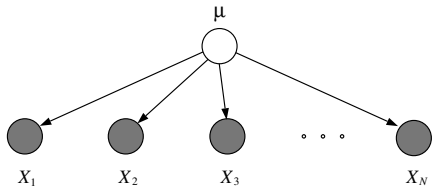
- Probability: inferring probabilistic quantities for data given fixed models (e.g. marginals, conditionals, log likelihood).
- Statistics: inferring a model given fixed data observations (e.g. clustering, classification, regression).
- Many approaches to statistics.
  We have focused on *maximum likelihood*.



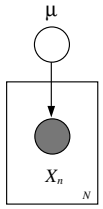


Plates

- Since Bayesian methods treat parameters as random variables, we would like to include them into the graphical model.
- One way to do this is to repeat all the iid observations explicitly and show the parameter only once.
- A better way is to use "plates", in which repeated quantities that are iid are put in a box.



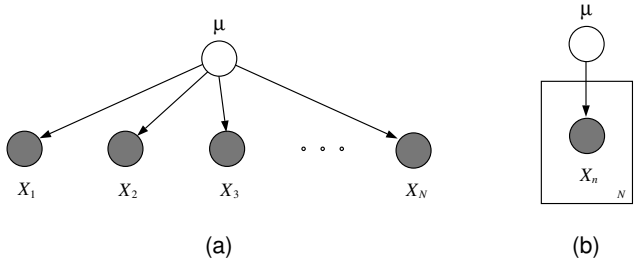
(a)



(b)

Plates

- Plates are like "macros" that allow you to draw a very complicated graphical model with a simpler notation.
- The rules of plates are simple: repeat every structure in a box a number of times given by the integer in the corner of the box (e.g. N), updating the plate index variable (e.g. n) as you go.
- Duplicate every arrow going into the plate and every arrow leaving the plate by connecting the arrows to each copy of the structure.

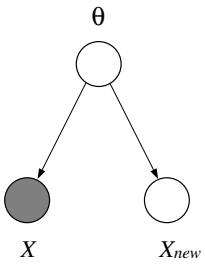


(a) (b)

Model Averaging

- Posterior $p(\theta|\mathcal{D})$ is used in all future Bayesian computations.
- For example, to do prediction of a new value $\mathbf{x}_{\text{new}}$ given some iid data, we compute the conditional posterior:

$$\begin{aligned} p(\mathbf{x}_{\text{new}}|\mathbf{X}) &= \int p(\mathbf{x}_{\text{new}}, \theta|\mathbf{X}) d\theta \\ &= \int p(\mathbf{x}_{\text{new}}|\theta, \mathbf{X}) p(\theta|\mathbf{X}) d\theta \\ &= \int p(\mathbf{x}_{\text{new}}|\theta) p(\theta|\mathbf{X}) d\theta \end{aligned}$$



- This means the Bayesian prediction is based on averaging predictions from lots of models, weighted by the posterior probability of the model's parameters.

Posterior Over Parameters

- If θ is a random variable, we can view the likelihood as a conditional probability and use Bayes rule:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}$$

- This crucial equation can be written in words:

$$\textit{posterior} \propto \textit{likelihood} \times \textit{prior}$$

- Computing the posterior requires conditioning on the data and having a *prior* over parameters.
- In contrast, frequentists consider various "estimators" of θ and hope to show that they have desirable properties, e.g. ML, "unbiased", "minimum variance", etc.

Multiple Model Averaging

- Imagine that we wanted to compute the probability of some new data (e.g. the density of a new point) taking into account the predictions of *all models*. We can compute:

$$\begin{aligned} p(\mathbf{x}_{\text{new}}|\mathbf{X}) &= \int\int p(\mathbf{x}_{\text{new}}, \theta, m|\mathbf{X}) d\theta dm \\ &= \int\int p(\mathbf{x}_{\text{new}}|\theta, m, \mathbf{X}) p(\theta, m|\mathbf{X}) d\theta dm \\ &= \int\int p(\mathbf{x}_{\text{new}}|\theta, m, \mathbf{X}) p(\theta|m, \mathbf{X}) p(m|\mathbf{X}) d\theta dm \end{aligned}$$

- This requires two posteriors, $p(m|\mathbf{X})$ (see later) and $p(\theta|\mathbf{x}, m)$.
- Remember: maximum likelihood alone cannot be used to do either model selection or model averaging since it always is subject to overfitting. Bayesian methods in principle can never overfit, since we integrate over all unknown quantities.

ML vs. Maximum-A-Posteriori (MAP)

- If we forced a Bayesian to pick a *single* value for the parameters rather use the entire posterior $p(\theta|\mathcal{D})$, what would they do?
- *Bayes point* (mean of posterior):

$$\theta_{Bayes} = \int \theta p(\theta|\mathcal{D}) d\theta$$

- *MAP* (mode of posterior):

$$\begin{aligned}\theta_{MAP} &= \text{argmax}_\theta \;\; p(\theta|\mathcal{D}) \\ &= \text{argmax}_\theta \;\; \log p(\mathcal{D}|\theta) + \log p(\theta)\end{aligned}$$

- The *maximum a-posteriori* (MAP) estimate looks exactly maxmimum likelihood except for an extra term which depends only on the parameters.
- This is often called "penalized maximum likelihood".

Example: Scalar Gaussian Model

- Consider a univariate Gaussian, with fixed, known variance σ^2.
- We want to put a prior $p(\mu)$ on the mean, μ and then compute its posterior, $p(\mu|\mathbf{X})$ using the Gaussian likelihood $p(\mathbf{X}|\mu)$.
- What should the prior be? Try another Gaussian:

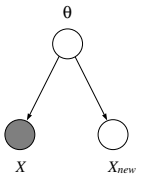
$$p(\mu) = \frac{1}{2\pi\tau^2} \exp\left\{-\frac{1}{2\tau^2}(\mu - \mu_0)^2\right\} \;\; = \mathcal{N}(\mu|\mu_0, \tau)$$

- Now the joint probability can be written as:

$$\begin{aligned}p(\mathbf{X}, \mu) &= p(\mathbf{X}|\mu)p(\mu) \\ &= \frac{1}{2\pi\sigma^2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\right\} \frac{1}{2\pi\tau^2} \exp\left\{-\frac{1}{2\tau^2}(\mu - \mu_0)^2\right\}\end{aligned}$$

- We need to marginalize this joint with respect to μ to obtain the posterior $p(\mu|\mathbf{X})$. This normalization can be done using the conditional Gaussian formulas or by explicitly completing the square.

Integrate or Optimize?



- Normally, Bayesian statistics needs to perform an integral in order to do predictions.
  Frequentist statistics uses a "plug-in" estimator such as ML.
- We can be "pseudo-Bayesian" by using single estimators such as Bayes-point or MAP.
- Notice that both the Bayesian approach and the maximum likelihood frequentist approach need to calculate the likelihood function $p(\mathbf{x}|\theta)$.
- So all the work we have done so far to is applicable to both Bayesian and ML frameworks.

Scalar Gaussian: Posterior over the Mean

- Amazingly, the posterior is another Gaussian:

$$\begin{aligned}p(\mu|\mathbf{X}) &= \frac{p(\mathbf{X}|\mu)p(\mu)}{\int p(\mathbf{X}, \mu) d\mu} \\ &= \frac{1}{2\pi s^2} \exp\left\{-\frac{1}{2s^2}(\mu - m)^2\right\} \\ &= \mathcal{N}(\mu|m, s^2) \\ m &= \frac{N/\sigma^2}{N/\sigma^2 + 1/\tau^2}\mu_{ML} + \frac{1/\tau^2}{N/\sigma^2 + 1/\tau^2}\mu_0 \\ s^2 &= \left(\frac{N}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}\end{aligned}$$

where μ_{ML} is the sample mean.

Conjugate Priors

- In the example we just worked out, the posterior had the same form as the prior (both were Gaussian).
- When this happens, the prior is called the *conjugate prior* for the parameters with respect to the *likelihood function*.
- Conjugate priors are very nice to work with because the posterior and prior have the same parameter types and the effect of the data is just to update the parameters from the prior to the posterior.
- In these settings, the prior can often be interpreted as some "pseudo-data" which we observed before we saw the real data.

Multinomial Posterior

- The posterior is also of the form of a Dirichlet:

$$p(\theta|\mathbf{X}) \propto p(\mathbf{X}|\theta)p(\theta) = \prod_k \theta_k^{\sum_n [x_n=k]} \theta_k^{\alpha_k - 1}$$
$$= \prod_k \theta_k^{\alpha_k - 1 + \sum_n [x_n=k]}$$

  which has parameters $\alpha'_k = \alpha_k + \sum_n [x_n = k]$.
- We see that to update the prior into a posterior, we simply add the observed counts to the priors.
- So we can think of the priors as "pseudo-counts".

Example: Multinomial

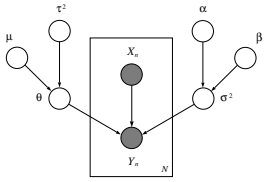
- Bayesian methods can also be used to estimate the density of discrete quantities (e.g. spam/nospam, shoe colour).
- If we use a *multinomial* distribution over K settings as the likelihood model, the *conjugate prior* is called the *Dirichlet distribution* defined as:

$$p(\theta) = C(\alpha)\theta_1^{\alpha_1 - 1}\theta_2^{\alpha_2 - 1} \ldots \theta_K^{\alpha_K - 1}$$
$$C(\alpha) = \Gamma(\sum_k \alpha_k) / \prod_k \Gamma(\alpha_k)$$

  where $\Gamma(\cdot)$ is the gamma function and the $\alpha - 1$ is a convention.
- This is a funny density, because it is a density over the *simplex*, i.e. over vectors whose components are non-negative and sum to one.
- In the binary case, the multinomial becomes a *binomial* $p(x|\theta) = \theta^x(1-\theta)^{1-x}$ and the Dirichlet becomes a *beta* distribution $p(\theta) = C(\alpha)\theta^{\alpha_1 - 1}(1-\theta)^{\alpha_2 - 1}$.

Heirarchical Bayes and Structure Learning

- What about the parameters of the parameter priors?
  In a full Bayesian formulation, they also have priors, called *hyperpriors* and we treat them in the same way.
- In theory we should do this upwards forever, but in practice we usually stop after only one or two levels.



- What about the model structure?
  In a full Bayesian formulation we also have a prior on that, and attempt to get a posterior. Sometimes this can be done (e.g. fully observed tree learning was just maximum likelihood over structure).

Model Selection

- In principle we could do model structure learning in a Bayesian way also. Consider a fixed class of models, indexed by $m = 1 \ldots M$ (e.g. Gaussian mixtures with m components).
- Since m is unknown, the Bayesian way is to treat it as a random variable and to compute its posterior:

$$p(m|\mathbf{X}) = \frac{p(\mathbf{X}|m)p(m)}{p(\mathbf{X})}$$

- Notice that we require a prior $p(m)$ on models as well as the *marginal likelihood*:

$$p(\mathbf{X}|m) = \int p(\mathbf{X}, \theta|m)d\theta = \int p(\mathbf{X}|\theta, m)p(\theta|m)d\theta$$

- We could try to compute the model with the highest posterior (in which case we don't have to compute $p(\mathbf{X})$.
- Or else we could use all of the models, *weighted by their posteriors* to do predictions at test time. This was called "model averaging".

Practical Bayesian and Other Methods

- Often these integrals are intractable, and so we resort to approximations, such as sampling, variational methods, large sample approximations such as BIC, etc.
- There are also frequentist (non-Bayesian) methods for model selection, as well as non-parametric approaches to solving estimation problems.