Lecture 12:

Bayesian Parameter Estimation

Sam Roweis

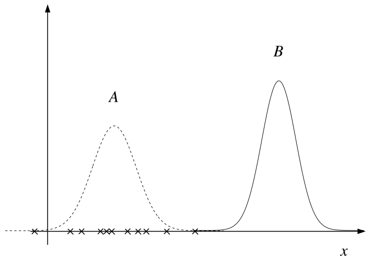
February 25, 2002

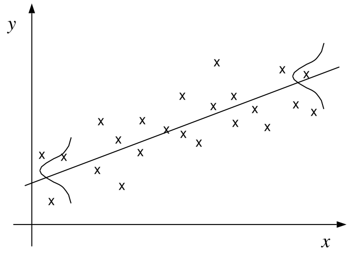
Bayesian Approach

- 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 variables. 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 $p(\theta|\mathcal{D})$.

Probability vs. Statistics

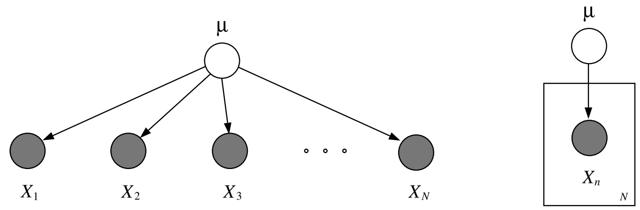
- Probability: inferring probabilistic quantities for data given fixed models (e.g. marginals, conditionals, log likelihood).
- Statistics: inferring a model given fixed data observations (e.g. clustering, classification, regression).
- Many approaches to statistics.
  We have focused on *maximum likelihood*.





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)

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:

posterior $\propto$ *likelihood* $\times$ *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.

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):

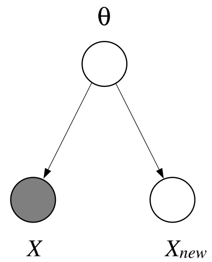
$$\begin{aligned}\theta_{MAP} &= \mathrm{argmax}_\theta \;\; p(\theta|\mathcal{D}) \\ &= \mathrm{argmax}_\theta \;\; \log p(\mathcal{D}|\theta) + \log p(\theta)\end{aligned}$$

- 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".

Model Averaging

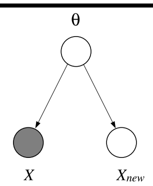
- Posterior $p(\theta|\mathcal{D})$ is used in all future Bayesian computations.
- For example, to do prediction of a new value $\mathbf{x}_{\mathrm{new}}$ given some iid data, we compute the conditional posterior:

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



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

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.

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:

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

- Now the joint probability can be written as:

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

- 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.

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.

Scalar Gaussian: Posterior over the Mean

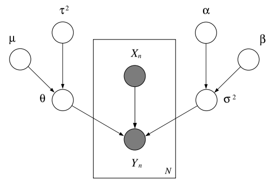
- Amazingly, the posterior is another Gaussian:

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

where μ_{ML} is the sample mean.

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).