Stat 305A: Linear Models **Autumn 2019/20**

Lecture 16: Basics of Bayes

Lecturer: Art B. Owen *November 19*

16.1 Bayes rule

Bayes rule for events A and B is

$$P(A\,|\,B) = \frac{P(A,B)}{P(B)} = \frac{P(A)P(B\,|\,A)}{P(B)}.$$

For continuous random variables x and y using events $x\,\mathrm{d}x$ and $y\,\mathrm{d}y$ we get

$$p(y\,|\,x) = \frac{p(y,x)}{p(x)} = \frac{p(y)p(x\,|\,y)}{p(x)}.$$

In our case y will be what we want to learn, such as all of the parameters in a model and x will be what we have, such as all of the data (called y!). So we use

$$p(\theta\,|\,y) = \frac{p(\theta,y)}{p(y)} = \frac{p(\theta)p(y\,|\,\theta)}{p(y)}.$$

In this new viewpoint, θ is a random variable just like y (which we use as shorthand for $y_1, y_2, \ldots, y_n$ or whatever else we have observed.) They have a joint distribution. Once we have seen y it is no longer random to us, and we can condition on its observed value. The parameter θ remains random but knowing y we think instead of $p(\theta\,|\,y)$. We might know $p(\theta)$ prior to seeing y so it is the ***prior distribution*** of θ and then $p(\theta\,|\,y)$ is the ***posterior distribution*** of θ (given y). If we should later get more data y' then $p(\theta\,|\,y)$ would become the new prior and $p(\theta\,|\,y,y')$ would be the updated posterior distribution of θ.

For real valued θ we can get a ***posterior credible interval*** $[L,U]$ computed so that

$$\Pr(L \leqslant \theta \leqslant U\,|\,y) = 0.99$$

or whatever other level we choose. Unlike a confidence interval here it is θ that is random (from $\Pr(\cdot)$) and both L and U are nonrandom. They ordinarily depend on y but conditionally on y they are not random.

16.2 The prior distribution

We get some very powerful results from Bayes. We might actually want to know $\Pr(H_0\,|\,y)$ even more than we want to know a p-value like $\Pr(t(Y) \geqslant t_{obs}\,|\,H_0)$. In order to get that of course we have to choose some $p(\theta)$.

There are numerous ways to choose this prior:

1. $p(\theta)$ may be a distribution that describes our subjective belief about θ. In principle, beliefs can be induced from looking at the bets we might take or reject on θ, though actually doing that would be cumbersome.
2. The problem we are considering might be just one in an ensemble of similar ones that we have seen many times before and may see many times in the future. Then $p(\theta)$ could be defined in reference to that higher order distribution of problem instances.
3. $p(\theta)$ could be defined to match qualitative aspects of the problem, such as enforcing $\theta_1 \leqslant \theta_2 \leqslant \cdots \leqslant \theta_L$ for a vector $\theta \in \mathbb{R}^L$ or otherwise widely covering the known or likely domain of θ.
4. $p(\theta)$ could be chosen for computational convenience.
5. $p(\theta)$ could be chosen to make posterior credible intervals have approximately the desired coverage level when viewed as confidence intervals. This is called ***calibration***. If we are designing an algorithm for somebody to use based on Bayes, we might prefer one that is well calibrated.

Sometimes more than one of the above comes into a decision. Point 2 above is the least controversial version of Bayes. One could argue that the value of $\theta_{\text{San Mateo}}$ has nothing to do with $\theta_{\text{Santa Clara}}$ or θ_{Ventura}. However that starts to sound like the arguments that people originally had against averaging observations from one sample, when those values were all taken in somewhat different ways. Hierarchical models where, for instance, data from one county are drawn from a county-specific parameter and those county-specific parameters are themselves drawn from some other distribution have proved very useful. See the book by Gelman and Hill on hierarchical models.

16.3 Stein and De Finetti

Stein showed in 1955 that if your method is admissible then it is either Bayes or the limit of a sequence of Bayes methods. Informally, a good method is nearly Bayes for some prior.

De Finetti showed that if $y_1, \ldots, y_n$ are exchangeable (joint distribution invariant to permuting their order) then $y_i \overset{\text{iid}}{\sim} f(\cdot;\theta)$ for some f and some random θ.

16.4 Normal data

See Chapter 5 of the book by Peter Hoff for details about Bayes for normally distributed data. We saw there how conjugate priors make things easy. We saw how the posterior mean was $\bar{y}$ shrunk towards the prior mean. We saw that precision (inverse variance) updates additively. The conjugate prior for μ is normal and the one for σ^2 is inverse Gamma. We also saw how the likelihood can swamp the prior so that in the end the prior makes little difference. That requires that the prior not rule out some values by declaring them impossible.

16.5 Tradeoffs

Using Bayes you get more but you have to give more (specify the prior). Moving away from conjugate priors can take you into a place where computation becomes incredibly hard. If you can write out all your

knowledge in terms of distributions that generate all of the variables you see, then you can potentially use Bayes to combine the information. If θ is more complicated like θ_1 identifies one of 10 models we might want to use and the rest of θ has the parameters for the model given by θ_1, then posterior predictions give us a weighted average over all 10 models. This ***Bayesian model averaging*** can be beneficial.