Bayesian Approach to Machine Learning (2)

Rohit Budhiraja

Machine Learning for Wireless Communications (EE798L)

Jan. 29, 2024

Recap of last lecture and today's agenda

- Recap of last class
  - Started motivating Bayesian approach by taking a coin game example
  - Bayesian approach also allows us to incorporate our prior belief about the model
  - Bayesian approach, similar to regularization, can avoid complex models
- Today's agenda
  - Discuss Bayesian approach in detail
  - Reference: Chap 3 of FCML

Posterior distribution of r (recap)

- From Bayes' rule

$$p\left(r|y_N\right) = \frac{P\left(y_N|r\right)p\left(r\right)}{P\left(y_N\right)}$$

- First quantity is likelihood $P\left(y_N|r\right)$, and second quantity is prior distribution $p(r)$
- For coin toss example,

$$\begin{aligned} P\left(y_N|r\right) &= \left(\begin{array}{c} N \\ y_N \end{array}\right) r^{y_N}(1-r)^{N-y_N} \\ p\left(r\right) &= \frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)} r^{\alpha-1}(1-r)^{\beta-1} \end{aligned}$$

- Last quantity is $P\left(y_N\right)$, which is called marginal distribution of y_N
- Called so because it is computed by integrating r out of the joint density $p(y_N, r)$

$$P\left(y_N\right) = \int_{r=0}^{r=1} p\left(y_N, r\right) dr$$

- $P\left(y_N\right)$, acts as a normalising constant to ensure that $p(r|y_N)$ is a properly defined density
- Marginal likelihood, in all but a small minority of cases, it is very difficult to calculate

Conjugate priors

- Before we calculate posterior $p(r|y_N)$ for our coin toss example using Bayes' rule

$$p(r|y_N) = \frac{P(y_N|r)\, p(r)}{P(y_N)}$$

- We discuss about conjugate likelihood-prior pair
- Likelihood-prior pair is said to be conjugate
  - If they result in a posterior which is of the same form as the prior, and is mathematically convenient
  - Enables us to compute posterior density analytically without worrying about computing $P(y_N)$
- Common conjugate pairs
  - Prior Likelihood
  - Gaussian Gaussian
  - Beta Binomial
  - Gamma Gaussian
  - Dirichlet Multinomial
- For binomial likelihood, we will obviously pick Beta prior

Posterior distribution (1)

- Returning to our example, we can omit $P(y_N)$. Posterior distribution

$$p\left(r|y_N\right) \propto P\left(y_N|r\right)p\left(r\right)$$

- Replacing the terms on the right hand side with a binomial and beta distribution gives

$$p\left(r|y_N\right) \propto \left[\begin{pmatrix} N \\ y_N \end{pmatrix} r^{y_N}(1-r)^{N-y_N}\right] \times \left[\frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)} r^{\alpha-1}(1-r)^{\beta-1}\right]$$

- A prior and likelihood are conjugate, we know that $p(r|yN)$ has to be a beta density
- Beta density, with parameters δ and γ has the following general form:

$$p\left(r\right) = Kr^{\delta-1}(1-r)^{\gamma-1},$$

  where K is a constant
- If we can arrange all the terms, including r, on RHS of equation into that looks like $r^{\delta-1}(1-r)^{\gamma-1}$
  - We already know the marginal likelihood $P(y_N)$ (normalising constant)
  - Must be $\frac{\Gamma(\delta+\gamma)}{\Gamma(\delta)\Gamma(\gamma)}$ because we already know that posterior has beta density
- Since we know $P(y_N)$, we do not need to compute it

Posterior distribution (2)

- Rearranging above equation gives us

$$\begin{aligned} p(r|y_N) &\propto \left[\begin{pmatrix} N \\ y_N \end{pmatrix} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right] \times \left[r^{y_N} r^{\alpha-1}(1-r)^{N-y_N}(1-r)^{\beta-1}\right] \\ &\propto r^{y_N+\alpha-1}(1-r)^{N-y_N+\beta-1} \\ &\propto r^{\delta-1}(1-r)^{\gamma-1} \end{aligned}$$

  where $\delta = y_N + \alpha$ and $\gamma = N - y_N + \beta$.
- We now have

$$\begin{aligned} p(r|y_N) &= \frac{\Gamma(\delta+\gamma)}{\Gamma(\delta)\Gamma(\gamma)} r^{\delta-1}(1-r)^{\gamma-1} \\ &= \frac{\Gamma(\alpha+\beta+N)}{\Gamma(\alpha+y_N)\Gamma(\beta+N-y_N)} r^{y_N+\alpha-1}(1-r)^{N-y_N+\beta-1} \end{aligned}$$

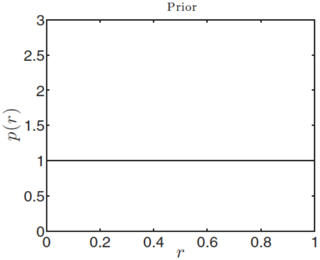
- Notice posterior parameters are computed by adding
  - number of heads (y_N) to first prior parameter (α) and number of tails ($N - y_N$) to second (β)
- This allows us to gain some intuition about prior parameters α and β
  - Can be thought of as the number of heads and tails in $\alpha + \beta$ previous hypothetical tosses

Posterior distribution (3)

- Fair coin $\alpha = \beta = 50$: equivalent to tossing a coin 100 times and obtaining 50 heads and 50 tails
- Biased scenario $\alpha = 5, \beta = 1$: equivalent to 6 tosses and obtaining 5 heads
- Analogy of α and β is not perfect as
  - α and β don't have to be integers and can be less than 1 (0.3 heads doesn't make much sense)

Posterior evaluation for first prior (1)

- Investigate evolution of posterior distribution for beta prior with $\alpha = 1, \beta = 1$ (no prior knowledge)



(a) $\alpha = 1, \beta = 1$

- We first compute mean and variance of R under the prior $\mathcal{B}(\alpha, \beta)$
  - Used for comparing different steps in posterior evolution
- Mean of beta-distributed r.v. is $\mathbf{E}_{p(r)}\{R\} = \frac{\alpha}{\alpha+\beta}$ (Tut. problem)
  - For $\alpha = \beta = 1$, $\mathbf{E}_{p(r)}\{R\} = \frac{1}{2}$
- Variance of a beta-distributed r.v. is $\mathbf{var}\{R\} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ (Tut. problem)
  - For $\alpha = \beta = 1$, $\mathbf{var}\{R\} = \frac{1}{12}$

Posterior evaluation for first prior (2)

- Our posterior is

$$
\begin{aligned}
p(r|y_N) &= \frac{\Gamma(\alpha+\beta+N)}{\Gamma(\alpha+y_N)\,\Gamma(\beta+N-y_N)} r^{y_N+\alpha-1}(1-r)^{N-y_N+\beta-1} \\
&= \mathcal{B}(\delta,\gamma)\,, \text{ with parameters } \delta=\alpha+y_N \text{ and } \gamma=\beta+N-y_N
\end{aligned}
$$

- Mean of R under posterior $p(r|y_N)=\mathcal{B}(\delta,\gamma)$

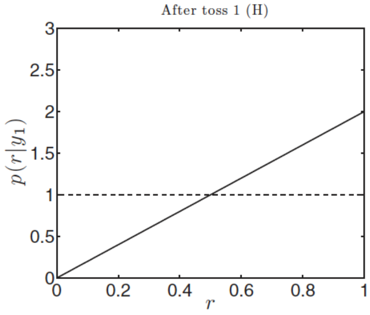
$$
\mathbf{E}_{p(r|y_N)}\{R\} = \frac{\delta}{\delta+\gamma}
$$

- Illustrate evolution of posterior – we will look at how it changes for every toss
- New customer hands over Re 1 and stall owner starts tossing the coin – first toss results in a head
- Posterior distribution after one toss is $p(r|y_N)=\mathcal{B}(\delta,\gamma)$
  - With $\alpha=\beta=1$, $N=1$ toss and $y_N=1$ head:

$$
\begin{aligned}
\delta &= \alpha+y_N = 1+1 = 2 \\
\gamma &= \beta+N-y_N = 1+1-1 = 1
\end{aligned}
$$

Posterior evaluation for first prior after first coin toss (1)

- Posterior distribution is shown as solid line and prior is shown as a dashed line



(b) $\delta = 2, \gamma = 1$

- Single observation has had quite a large effect – posterior is very different from prior
  - All values of r were equally likely in prior
  - This has now changed higher values are more likely than lower values with zero density at $r = 0$
- Consistent with evidence
  - observing one head makes high values of r slightly more likely and low values slightly less likely
- Posterior is still very broad, as we have observed only one toss

Posterior evaluation for first prior after first coin toss (2)

- Mean of R under posterior $p(r|y_N) = \mathcal{B}(\delta, \gamma)$ with $\delta = 2$ and $\gamma = 1$ is

$$\mathbf{E}_{p(r|y_N)}\{R\} = \frac{\delta}{\delta + \gamma} = \frac{2}{3}$$

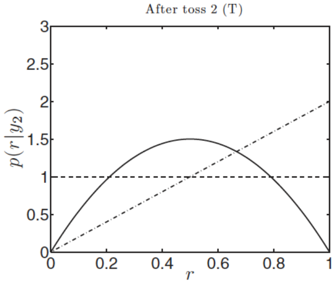
- Observing a solitary head has increased expected value of r from $1/2$ to $2/3$
  - Increase in expected value tells us that heads are slightly more likely than tails
- Variance of posterior is

$$\mathbf{var}\{R\} = \frac{\delta\gamma}{(\delta+\gamma)^2(\delta+\gamma+1)} = \frac{1}{18}$$

- Lower than prior variance of $1/12$
  - Reduction in variance tells us that we have less uncertainty about the value of r than we did
- Stall owner tosses the second coin $(N = 2)$ and it lands tails. With one head $(y_N = 1)$ and one tail

$$\begin{aligned} \delta &= \alpha + y_N = 1 + 1 = 2 \\ \gamma &= \beta + N - y_N = 1 + 2 - 1 = 2 \end{aligned}$$

Posterior evaluation for first prior after two coin tosses



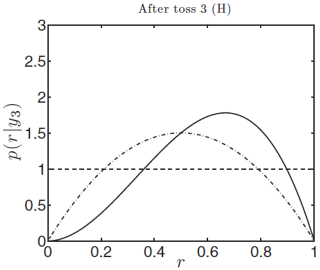
(c) $\delta = 2, \gamma = 2$

- Posterior is now curved rather than straight – observing a tail has made lower values more likely
- Expected value and variance are now $\mathbf{E}_{p(r|y_N)}\{R\} = \frac{1}{2}$, $\textbf{var}\{R\} = \frac{1}{20}$
- Expected value has decreased back to $1/2$, which is same as under the prior
  - We might conclude that we haven't learnt anything
- Variance has decreased (from $1/18$ to $1/20$) – less uncertainty in r and have learnt something
  - In fact, we've learnt that r is closer to $1/2$ than we assumed under the prior

Posterior evaluation for first prior after three coin tosses

- Third toss results in another head. We have $N = 3$ tosses, $y_N = 2$ heads and $N - y_N = 1$ tail:

$$\delta = \alpha + y_N = 1 + 2 = 3 \text{ and } \gamma = \beta + N - y_N = 1 + 3 - 2 = 2$$

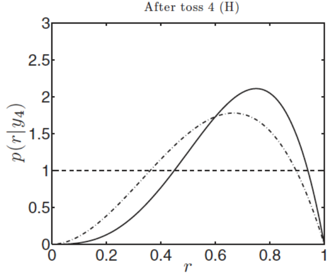


(d) $\delta = 3, \gamma = 2$

- Observe that second head skews density to right, suggesting that heads are more likely than tails
- Entirely consistent with the evidence – we have seen more heads than tails
- We have only seen three coins though, so there is still a high level of uncertainty
- Density suggests that r could potentially still be pretty much any value between 0 and 1
- Expected value and variance are now $\mathbf{E}_{p(r|y_N)}\{R\} = \frac{3}{5}$, $\mathbf{var}\{R\} = \frac{1}{25}$

Posterior evaluation for first prior after four coin tosses

- Toss 4 also comes up heads ($y_N = 3, N = 4$), resulting in $\delta = 1 + 3 = 4$ and $\gamma = 1 + 4 - 3 = 2$

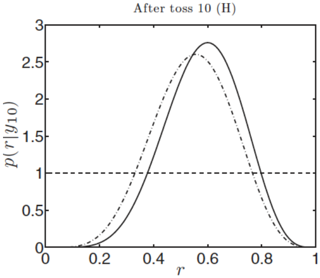


(e) $\delta = 4, \gamma = 2$

- Posterior is once again skewed to right – we've now seen 3 H and 1 T so it is likely that $r > 1/2$
- Notice the difference between $N = 3$ posterior and $N = 4$ posterior
  - For very low values of r extra head has left us pretty convinced that r is not 0.1 or lower
- Expected value and variance are now $\mathbf{E}_{p(r|y_N)}\{R\} = \frac{2}{3}$, $\mathbf{var}\{R\} = \frac{2}{63} = 0.0317$
  - Expected value has increased and the variance has once again decreased

Posterior evaluation for first prior after ten coin tosses

- Remaining six tosses are made so that complete sequence is $H, T, H, H, H, H, T, T, T, H$
- Posterior distribution after $N = 10$ tosses, with six heads and four tails i.e., $(y_N = 6)$
  - Has parameters $\delta = 1 + 6 = 7$ and $\gamma = 1 + 10 - 6 = 5$
- Expected value and variance are $\mathbf{E}_{p(r|y_N)}\{R\} = \frac{7}{12} = 0.5833, \mathbf{var}\{R\} = 0.0187$



(f) $\delta = 7, \gamma = 5$

- Ten observations have increased expected value from 0.5 to 0.5833 and decreased variance from $1/12 = 0.0833$ to 0.0187
- We see from figure that we can also be pretty sure that $r > 0.2$ and $r < 0.9$
- Uncertainty in value of r is still quite high because we have only observed ten tosses