Bayesian Approach to Machine Learning (3)

Rohit Budhiraja

Machine Learning for Wireless Communications (EE798L)

Jan. 31, 2024

Recap of last lecture and today's agenda

- Recap of last class
  - Discussed Bayesian approach - calculated posterior for coin toss example
  - Discussed idea of conjugate likelihood and prior, and calculated posterior
  - Discussed posterior evolution
- Today's agenda
  - Continue discussing Bayesian approach

Bayesian treatment for coin tossing game (recap)

- Considered a coin tossing game and modelled data using binomial distribution with likelihood

$$P\left(Y = y | r, N\right) = \begin{pmatrix} N \\ y \end{pmatrix} r^y (1-r)^{N-y}$$

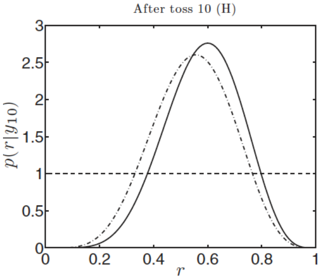
- Treated r as a parameter and calculated its maximum likelihood (ML) point estimate: $\hat{r} = y/N$
- Calculated expected winning probability with ML estimate $\hat{r}$
  - ML estimate $\hat{r} = 0.9$ computed based on ten tosses $H, T, H, H, H, H, H, H, H, H$
- Considering r as a random variable will help in measuring and understanding this uncertainty
- By defining random variable Y_N as number of heads obtained in N tosses, we calculated $p\left(r|y_N\right)$

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

- With posterior, we will recompute winning probability

Posterior calculation for first prior after ten tosses (recap)

- Complete toss 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., $(y_N = 6)$



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

- Posterior distribution $p\left(r|y_N\right) = \mathcal{B}\left(\delta, \gamma\right)$ with parameters $\delta = 1 + 6 = 7$ and $\gamma = 1 + 10 - 6 = 5$
  - Considered prior $p(r) = \mathcal{B}\left(\alpha, \beta\right)$ with $\alpha = \beta = 1$
- Posterior distribution $p\left(r|y_N\right) = \mathcal{B}\left(\delta, \gamma\right)$ contains all the information about r
- Posterior can be used to calculate different point estimates of r

Different point estimates of r

- Posterior distribution of $p\left(r|y_N\right)$

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

- Our posterior distribution $p\left(r|y_N\right) = \mathcal{B}\left(\delta, \gamma\right)$ with parameters $\delta = 1 + 6 = 7$ and $\gamma = 1 + 10 - 6 = 5$
- Maximum a posterior (MAP) estimate: maximum value of posterior distribution $\mathcal{B}\left(\delta, \gamma\right)$, which is
  - Its mode i.e., $\hat{r} = \frac{\delta - 1}{\delta + \gamma - 2} = 6/10$
- Mean estimate : mean of posterionr distribution $\mathcal{B}\left(\delta, \gamma\right)$, which is
  - Mean: $\hat{r} = \frac{\delta}{\delta + \gamma} = \frac{7}{12}$, denoted as $\hat{r} = \mathbb{E}_{p(r|y_N)}\{R\}$
- Both MAP and mean estimator are a posterior estimate – which one to pick?
- Maximum likelihood (ML) estimate $\hat{r}$
  - MAP estimate with prior $p(r)$ set to uniform distribution $\rightarrow$ above MAP estimate is also ML estimate
  - Matches with our earlier ML estimate: $\hat{r} = y/N = 6/10$

Winning probability using point estimate

- Posterior distribution $p(r|y_N)$ contains all the information we have about r
  - We will use it to compute expected winning probability
- Before we do so, we will first use mean estimate $\hat{r} = \frac{\delta}{\delta+\gamma} = \frac{7}{12}$ to compute winning probability
- Compute winning probability $P(Y_{new} \leq 6|\hat{r})$
  - Differentiate observed and future tosses by using Y_{new} as a r.v. to describe future ten tosses
- Probability of winning the game is

$$\begin{aligned} P(Y_{new} \leq 6|\hat{r}) &= 1 - \sum_{y_{new}=7}^{y_{new}=10} P(Y_{new} = y_{new}|\hat{r}) = 1 - \sum_{y_{new}=7}^{y_{new}=10} \begin{pmatrix} N \\ y_{new} \end{pmatrix} (\hat{r})^{y_{new}} (1-\hat{r})^{N-y_{new}} \\ &= 1 - 0.3414 = 0.6586 \end{aligned}$$

- Suggesting that we will win more often than lose

Expected winning probability using complete posterior (1)

- Compute expected winning probability using all of posterior information. This requires computing

$$\int_{r=0}^{r=1} P\left(Y_{new} \leq 6|r\right) p\left(r|y_N\right) dr = \mathbb{E}_{p(r|y_N)}\{P\left(Y_{new} \leq 6|r\right)\}$$

$$\mathbb{E}_{p(r|y_N)}\{P\left(Y_{new} \leq 6|r\right)\} = \sum_{y_{new}=1}^{y_{new}=6} \mathbb{E}_{p(r|y_N)}\{P\left(Y_{new} = y_{new}|r\right)\}$$

- We need to compute $\mathbb{E}_{p(r|y_N)}\{P\left(Y_{new} = y_{new}|r\right)\}$ which is

$$\begin{aligned}
\mathbb{E}_{p(r|y_N)}\{P\left(Y_{new} = y_{new}|r\right)\} &= \int_{r=0}^{r=1} P\left(Y_{new} = y_{new}|r\right) p\left(r|y_N\right) dr \\
&= \int_{r=0}^{r=1} \left[\binom{N_{new}}{y_{new}} r^{y_{new}} (1-r)^{N_{new}-y_{new}}\right] \left[\frac{\Gamma(\delta+\gamma)}{\Gamma(\delta)\Gamma(\gamma)} r^{\delta-1} (1-r)^{\gamma-1}\right] dr \\
&= \binom{N_{new}}{y_{new}} \frac{\Gamma(\delta+\gamma)}{\Gamma(\delta)\Gamma(\gamma)} \int_{r=0}^{r=1} r^{y_{new}+\delta-1} (1-r)^{N_{new}-y_{new}+\gamma-1} dr
\end{aligned}$$

- Argument inside the integral is an unnormalised beta density with parameters $y_{new}+\delta = \delta'$ and $N_{new} - y_{new} + \gamma = \gamma'$

Expected winning probability using complete posterior (2)

- For a beta density with parameters δ' and γ', following must be true:

$$\int_{r=0}^{r=1} \frac{\Gamma\left(\delta'+\gamma'\right)}{\Gamma\left(\delta'\right)\Gamma\left(\gamma'\right)} r^{\delta'-1}(1-r)^{\gamma'-1} dr = 1,$$

$$\int_{r=0}^{r=1} r^{\delta'-1}(1-r)^{\gamma'-1} dr = \frac{\Gamma\left(\delta'\right)\Gamma\left(\gamma'\right)}{\Gamma\left(\delta'+\gamma'\right)}$$

- Our desired expectation becomes

$$\mathbb{E}_{p(r|y_N)}\{P\left(Y_{new}=y_{new}|r\right)\} = \begin{pmatrix} N_{new} \\ y_{new} \end{pmatrix} \frac{\Gamma\left(\delta+\gamma\right)}{\Gamma\left(\delta\right)\Gamma\left(\gamma\right)} \frac{\Gamma\left(\delta+y_{new}\right)\Gamma\left(\gamma+N_{new}-y_{new}\right)}{\Gamma\left(\delta+\gamma+N_{new}\right)}$$

- After ten tosses, we have $\delta=7, \gamma=5$. Expected winning probability:

$$\mathbb{E}_{p(r|y_N)}\{P\left(Y_{new}\leq 6|r\right)\} = \sum_{y_{new}=1}^{y_{new}=6} \mathbb{E}_{p(r|y_N)} P\left(Y_{new}=y_{new}|r\right) = 0.6055$$

- Expected winning probability and the one with point estimate predict we will win more often
- Agrees with evidence – one person we have fully observed got six heads and four tails and won Rs 2
- Point estimate gives a higher probability – ignoring posterior uncertainty makes it more likely that we will win

Three scenarios – summary

- For three different scenarios the expected probability of winning
  - No prior knowledge: $\mathbb{E}_{p(r|y_N)}\{P(Y_{new} \leq 6|r)\} = 0.6055$
  - Fair coin: $\mathbb{E}_{p(r|y_N)}\{P(Y_{new} \leq 6|r)\} = 0.7579$ (Similarly compute)
  - Biased coin: $\mathbb{E}_{p(r|y_N)}\{P(Y_{new} \leq 6|r)\} = 0.2915$ (Similarly compute)
- Which one should we choose? We could choose based on our prior beliefs
- Given that stall owner doesn't look like he will go out of business, scenario 3 might be sensible
- We might decide that we do not know anything about owner and coin and look to scenario 1
- We might believe that owner would never stoop to cheating and go for scenario 2
- It is possible to justify any of them but one thing is clear
  - Bayesian technique allows us to combine observed data with prior knowledge in a principled way
  - Posterior density models uncertainty that remains in r at each stage

Marginal Likelihoods

- Subjective beliefs are not the only option for determining which of our three scenarios is best
- Marginal likelihood $p(y_N)$ provide another method. It is related to r as follows

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

- When considering different choices of prior $p(r)$, it should be written as $p(r|\alpha, \beta)$
  - Density is a function of particular α and β values
- Extending this conditioning to earlier equation

$$P(y_N|\alpha, \beta) = \int_{r=0}^{r=1} P(y_N|r)\, p(r|\alpha, \beta)\, dr.$$

- Marginal likelihood $P(y_N|\alpha, \beta)$ is a very useful and important quantity
  - Tells us how likely the data (y_N) is, given our choice of prior parameters α and β
  - Higher $P(y_N|\alpha, \beta)$, better our data agrees with the prior specification
- We could use $P(y_N|\alpha, \beta)$ to help choose the best scenario:
  - Select the scenario for which $P(y_N|\alpha, \beta)$ is highest

Marginal Likelihoods

- To compute this quantity, we need to evaluate the following integral:

$$\begin{aligned}
P(y_N|\alpha,\beta) &= \int_{r=0}^{r=1} P(y_N|r)\, p(r|\alpha,\beta)\, dr \\
&= \int_{r=0}^{r=1} \begin{pmatrix} N \\ y_N \end{pmatrix} r^{y_N} (1-r)^{N-y_N} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r^{\alpha-1} (1-r)^{\beta-1}\, dr \\
&= \begin{pmatrix} N \\ y_N \end{pmatrix} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_{r=0}^{r=1} r^{\alpha+y_N-1} (1-r)^{\beta+N-y_N-1}\, dr.
\end{aligned}$$

- Argument inside the integral is an unnormalised beta density
  - Integrating it we will give inverse of normal beta normalising constant

$$P(y_N|\alpha,\beta) = \begin{pmatrix} N \\ y_N \end{pmatrix} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+y_N)\Gamma(\beta+N-y_N)}{\Gamma(\alpha+\beta+N)}$$

- We considered two sets of coin tosses: i) 9 heads and 1 tail; and ii) 6 heads and 5 tails
  - Total of 15 heads in 2 sets of 10 tosses $N = 20$ and $y_N = 14$
- We have three different possible pairs of α and β values. Plugging these values above
  - No prior knowledge, $\alpha = \beta = 1$, $p(y_N|\alpha,\beta) = 0.0476$
  - Fair coin, $\alpha = \beta = 50$, $p(y_N|\alpha,\beta) = 0.0441$
  - Biased coin: $\alpha = 5, \beta = 1$, $p(y_N|\alpha,\beta) = 0.0576$

Marginal Likelihoods

- Prior corresponding to biased coin has highest marginal likelihood and the fair coin prior has lowest
- Expected winning probability calculated earlier for this scenario was

$$\mathbb{E}_{p(r|y_N,\alpha,\beta)}\{P\left(Y_{new} \leq 6|r\right)\} = 0.2915$$

- A word of caution is required here
- Choosing priors in this way is essentially choosing the prior that best agrees with the data
- Prior no longer corresponds to our belief – may be unacceptable in some applications
- Marginal likelihood gives a single value that tells us how much data backs up prior beliefs
- In earlier example, data suggests that biased coin prior is best supported by the evidence
- Extend the prior comparison to using the marginal likelihood to calculate optimal value of α and β