Bayesian Approach to Machine Learning (1)

Rohit Budhiraja

Machine Learning for Wireless Communications (EE798L)

Jan 27, 2023

Summary and next agenda

- Summary till now
  - Generative modeling approach tells us how confident the model is about the predictions it is making
  - Maximum likelihood approach favors complex models
- Next agenda
  - Bayesian approach, similar to regularization, can avoid complex models[1]
  - Bayesian approach also allows us to incorporate our prior belief about the model

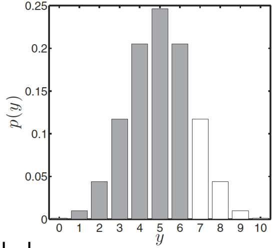
[1] Chap 3 of FCML

Coin game (1)

- Imagine you are walking and come across a stall where customers are playing a coin tossing game
- Stall owner tosses a coin ten times for each customer
  - If coin lands heads on six or fewer times, customer wins back their Rs. 1 stake plus an additional Rs 1
  - For seven or more, stall owner keeps their money
- Probability of y heads from N tosses where each toss lands heads with probability r is

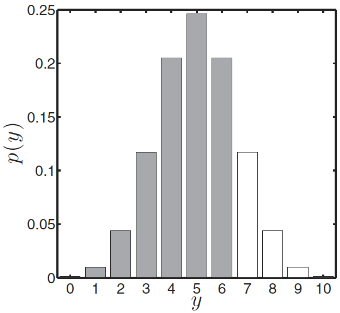
$$P(Y = y) = \binom{N}{y} r^y (1-r)^{N-y} \text{ (binomial distribution)}$$

- Assume coin is fair and therefore set $r = 0.5$. For $N = 10$ tosses, probability distribution function is



- Bars corresponding to $y \leq 6$ are shaded

Coin game (2)



- Probability that Y is less than or equal to 6, $P(Y \leq 6)$ when $N = 10$ and $r = 0.5$:

$$\begin{aligned} P(Y \leq 6) = 1 - P(Y > 6) &= 1 - [P(Y = 7) + P(Y = 8) + P(Y = 9) + P(Y = 10)] \\ &= 1 - [0.1172 + 0.0439 + 0.0098 + 0.0010] = 0.8281 \end{aligned}$$

- Seems like a pretty good game you'll double your money with probability 0.8281
- It is also possible to compute the expected return from playing the game

Coin game (3)

- Expected value of a function $f(X)$ of a random variable X is computed as

$$\mathbf{E}_{P(x)}\{f(X)\} = \sum_x f(x)\, P(x)$$

- Let X be a random variable that takes a value 1 if we win and a value 0 if we lose
- If we win, $(X = 1)$, we get a return of Rs 2 (our original stake plus an extra Re 1) so $f(1) = 2$
- If we lose, we get a return of nothing so $f(0) = 0$. Hence our expected return is

$$f(1)\, P(X = 1) + f(0)\, P(X = 0) = 2 \times P(Y \leq 6) + 0 \times P(Y > 6) = 1.6562$$

- Given that it costs Re 1 to play, you win, on average, $1.6562 - 1$ or $= 65.62$ paise per game
- If you played 100 times, you'd expect to walk away with a profit of Rs 65.62 – sensible to play
- While waiting you notice that stall owner is reasonably wealthy and very few customers seem to win
- Perhaps the assumptions underlying the calculations are wrong, which are
  - Number of heads can be modelled as a random variable with a binomial distribution
  - Coin is fair i.e., probability of heads is same as probability of tails $r = 0.5$

Coin game (4)

- Seems hard to reject binomial distribution
  - Events are taking place with only two possible outcomes and tosses do seem to be independent
- It leaves r, the probability that the coin lands heads
  - Assumed that coin was fair – maybe this is not the case?
- To investigate this, we use generative data modeling approach
  - Define a model which can generate data similar to which is given to us
- Our data in this case - there are three people in the queue to play
  - First one plays and gets the following sequence of heads and tails: $H, T, H, H, H, H, H, H, H, H$
- For generative date modeling approach, we need likelihood distribution, which is binomial here

$$P(Y = y|r) = \binom{N}{y} r^y (1-r)^{N-y}$$

  - Treat r as a parameter (like $\mathbf{w}$ and σ^2 earlier), and calculate its maximum likelihood estimate
- Taking the natural logarithm gives

$$L = \log P(Y = y|r) = \log \binom{N}{y} + y \log r + (N-y)\log(1-r)$$

Coin game (5)

- Log likelihood is
$$L = \log P(Y = y|r) = \log \begin{pmatrix} N \\ y \end{pmatrix} + y \log r + (N - y) \log(1 - r)$$
- Maximum likelihood estimate of r:
$$\begin{aligned} \frac{\partial L}{\partial r} &= \frac{y}{r} - \frac{N-y}{1-r} = 0 \\ y(1-r) &= r(N-y) \\ y &= rN \Rightarrow r = \frac{y}{N} \end{aligned}$$
- With $y = 9$ and $N = 10$ gives $r = 0.9$. Recalculated winning probability: $P(Y \leq 6|\, r) = 0.0128$
  - Recalculated winning probability with a point estimate of r
- Expected return is now $2 \times P(Y \leq 6) + 0 \times P(Y > 6) = 0.0256$
- Given that it costs Re 1 to play, we expect to make $0.0256 - 1 = -0.9744$ per game
  - A loss of ≈ 97 paise
- P$(Y \leq 6) = 0.0128$ suggests that only about 1 person in every 100 should win
  - But this does not seem to be reflected in the number of people who are winning
- Although evidence from this run of coin tosses suggests $r = 0.9$
  - It seems too biased given that several people have won

Bayesian Way

- Point estimate of r computed earlier was based only on data, and that too of just ten tosses
  - Data could be misleading
- Given the random nature of coin toss, if we observe several sequences of tosses
  - It is likely that we would get a different r each time
- 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 would want distribution of r conditioned on value of Y_N i.e., $p(r|y_N)$
  - Calculate posterior distribution of r, instead of its point estimate, while treating it as a fixed parameter

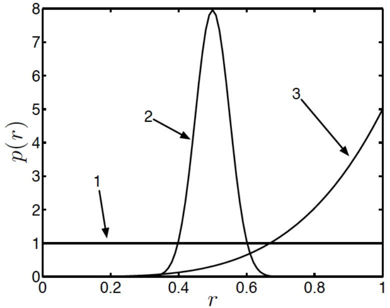
Bayesian Way - prior (1)

- To calculate posterior distribution $p(r|y_N)$, we use Bayes' rule

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

- $P(y_N|r)$ is likelihood distribution while $p(r)$ is prior distribution
  - $p(r)$ allows us to express any belief we have in value of r before we see any data
- To illustrate this, we shall consider the following three examples:
  - We do not know anything about tossing coins or the stall owner
  - We think the coin (and hence the stall owner) is fair
  - We think the coin (and hence the stall owner) is biased to give more heads than tails
- We can encode each of these beliefs as different prior distributions on r
  - Note that r can take any value between 0 and 1, must be modelled as a continuous random variable
  - Will next show three density functions that might be used to encode our three different prior beliefs

Bayesian Way - prior (2)



- Belief 1 is a uniform density between 0 and 1; shows no preference for any particular r value
- Belief 2 has density function that is concentrated around r $=$ 0.5, value we expect for a fair coin
  - Density suggests that we do not expect much variance in r; it's almost certainly between 0.4 and 0.6
  - Most coins that any of us have tossed agree with this
- Belief 3 encapsulates our belief that the coin (and therefore the stall owner) is biased
  - Density suggests that $r > 0.5$ and that there is a high level of variance
  - Our belief is just that the coin is biased – we don't really have any idea how biased at this stage
- We will not choose between our three scenarios at this stage
  - It is interesting to see the effect these different beliefs will have on $p(r|y_N)$

Bayesian Way - prior (3)

- Three functions are examples of beta probability density functions
  - Continuous random variables constrained to lie between 0 and 1; perfect for our example
- For a random variable R with parameters α and β, it is defined as

$$p(r) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r^{\alpha-1}(1-r)^{\beta-1}$$

- $\Gamma(a)$ is known as gamma function

$$\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_{r=0}^{r=1} r^{\alpha-1}(1-r)^{\beta-1} dr$$

- Ensures that density is normalized

$$\int_{r=0}^{r=1} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r^{\alpha-1}(1-r)^{\beta-1} dr = 1$$

- Two parameters α and β control the shape of resulting density function and must both be positive
  - Know nothing: $\alpha = 1$ and $\beta = 1$
  - Fair coin: $\alpha = 50$ and $\beta = 50$
  - Biased: $\alpha = 5$ and $\beta = 1$

Bayesian Way - prior (4)

- Problem of choosing these values is a big one
  - For example, why should we choose $\alpha = 5; \beta = 1$ for a biased coin? There is no easy answer to this
  - Show for beta distribution, they can be interpreted as a number of previous, hypothetical coin tosses

Marginal distribution

- From Bayes' rule

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

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

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

- $P(y_N)$, acts as a normalising constant to ensure that $p(r|y_N)$ is a properly defined density
- Joint density can be factorised to give

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

- Product of prior and likelihood integrated over the range of values that r may take
- $P(y_N)$ is also known as marginal likelihood
  - As it is likelihood of data y_N averaged over all parameter values
- See later that it can be useful quantity in model selection
  - Unfortunately, in all but a small minority of cases, it is very difficult to calculate