Introduction to Machine Learning

CS307 --- Fall 2022

Maximum Likelihood Estimation

Reading:
Sections 20.2-20.3, R&N

Maximum Likelihood Estimation

- From a Bayesian perspective, we are interested in finding the MAP hypothesis

$$
\begin{aligned}
h_{MAP} &= \arg\max_{h \in H} P(h|D) \\
&= \arg\max_{h \in H} \frac{P(D|h)P(h)}{P(D)} \\
&= \arg\max_{h \in H} P(D|h)P(h)
\end{aligned}
$$

- But in many cases we have to assume a uniform distribution over the hypotheses (e.g., because of lack of prior knowledge of the domain), effectively seeking the maximum likelihood (ML) hypothesis

$$h_{ML} = \arg\max_{h \in H} P(D|h)$$

Maximum Likelihood Estimates

- If the hypotheses are parameterized (by say θ), then seeking a ML hypothesis is equivalent to seeking values of θ that maximize data likelihood.

$$\theta^* = \arg\max_{\theta} P(D|\theta)$$

- A maximum likelihood estimate (MLE) is a parameter estimate that maximizes the data likelihood. It is an estimate that is most consistent with the data.

Example: Coin Tossing

- How likely am I to toss a head? Assume that a series of 10 trials/tosses yields (h,t,t,t,h,t,t,h,t,t)
  ($x1=3$, $x2=7$), $n = 10$

- Probability of tossing a head = 3/10

- That's a MLE! This estimate is absolutely consistent with the observed data.

- But … is this an estimate that maximizes data likelihood?

Maximizing Data Likelihood

- What's the data likelihood?

$$L(\theta) = P(D \mid \theta) = \theta^3 (1-\theta)^7$$

- How to maximize the data likelihood function?
  - Take the first derivative of the likelihood function with respect to the parameter theta and solve for 0. This value maximizes the likelihood function and is the MLE.

Maximizing the Likelihood

$$L(\theta) = P(D|\theta) = \theta^3 (1-\theta)^7$$

- It's usually easier to maximize the log likelihood. So let's maximize

$$\log L(\theta) = \log(\theta^3 (1-\theta)^7)$$

$$\log L(\theta) = \log \theta^3 + \log(1-\theta)^7$$

$$\log L(\theta) = 3 \log \theta + 7 \log(1-\theta)$$

- Take the derivative of the function and set it to zero.

$$\frac{d \log L(\theta)}{d\theta} = \frac{3}{\theta} - \frac{7}{1-\theta} = 0$$

- Solve for theta:

$$\theta = \frac{3}{10}$$

A General Scalar MLE Strategy

Task: Find MLE θ that maximizes P(Data | θ)

1. Write LL = log P(Data | θ)
2. Work out the first derivative of the likelihood function using high-school calculus
3. Set the derivative to zero, thus creating an equation in terms of θ
4. Solve it
5. Check that you've found a maximum rather than a minimum or a saddle point

A General MLE Strategy

Suppose $\theta = (\theta_1, \theta_2, \ldots, \theta_n)^T$ is a vector of parameters.

Task: Find MLE θ that maximizes $P(\text{Data} \mid \theta)$

1. Write $LL = \log P(\text{Data} \mid \theta)$
2. Work out the partial derivative of LL w.r.t. each θ_l
3. Solve the set of simultaneous equations

$$\frac{\partial LL}{\partial \theta_1} = 0 \qquad \frac{\partial LL}{\partial \theta_2} = 0 \qquad \ldots \qquad \frac{\partial LL}{\partial \theta_n} = 0$$

4. Check that you are at a maximum.

Does this strategy always work?

What if you cannot solve the simultaneous equations?

- use gradient ascent

Are there other problems?

An Example: Animal Classification

- There are n animals classified into one of four possible categories
  - Category counts are the sufficient statistics to estimate the parameters

- Techniques for finding MLEs is the same
  - Take derivative of likelihood function
  - Solve for zero

An Example: Animal Classification

There are n=197 animals classified into one of 4 categories:
Y = (y1, y2, y3, y4) = (125, 18, 20, 34)

The probability associated with each category is given as:

$$\Theta = (\frac{1}{2} + \frac{1}{4}\pi, \frac{1}{4}(1-\pi), \frac{1}{4}(1-\pi), \frac{1}{4}\pi)$$

The resulting likelihood function for this data is:

$$L(\pi) = \frac{n!}{y1!\,y2!\,y3!\,y4!}(\frac{1}{2} + \frac{1}{4}\pi)^{y1}(\frac{1}{4}(1-\pi))^{y2}(\frac{1}{4}(1-\pi))^{y3}(\frac{1}{4}\pi)^{y4}$$

Maximizing Log Likelihood

$$\log L(\pi) = y1 * \log(\frac{1}{2} + \frac{1}{4}\pi) + y2 * \log(\frac{1}{4}(1-\pi)) + y3 * \log(\frac{1}{4}(1-\pi))$$

$$+ y4 * \log(\frac{1}{4}\pi) + \log(\frac{n!}{y1!\, y2!\, y3!\, y4!})$$

$$\frac{d \log L(\pi)}{d\pi} = \frac{y1}{2+\pi} - \frac{y2+y3}{1-\pi} + \frac{y4}{\pi} = 0$$

$$\frac{d \log L(\pi)}{d\pi} = \frac{125}{2+\pi} - \frac{38}{1-\pi} + \frac{34}{\pi} = 0$$

$$\pi = 0.627$$

Adversity Strikes!

- What if the observed data is incomplete? What if there are really 5 categories?

- y1 is the composite of 2 categories (x1+x2)

$$p(y1) = \frac{1}{2} + \frac{1}{4}\pi,\ p(x1) = \frac{1}{2},\ p(x2) = \frac{1}{4}\pi$$

- How can we make a MLE, since we can't observe category counts x1 and x2?!
  - Unobserved sufficient statistics!?

The EM Algorithm

- E-STEP: Find the expected values of the sufficient statistics for the complete data X, given the incomplete data Y and the current parameter estimates
- M-STEP: Use those sufficient statistics to make a MLE as usual!
- Repeat the above steps until convergence

MLE for Complete Data

$$X = (x1, x2, x3, x4, x5) = (x1, x2, 18, 20, 34) \quad \text{where } x1+x2=125$$

$$\Theta = (\frac{1}{2}, \frac{1}{4}\pi, \frac{1}{4}(1-\pi), \frac{1}{4}(1-\pi), \frac{1}{4}\pi)$$

$$L(\pi) = \frac{n!}{x1!\,x2!\,x3!\,x4!\,x5!}(\frac{1}{2})^{x1}(\frac{1}{4}\pi)^{x2}(\frac{1}{4}(1-\pi))^{x3}(\frac{1}{4}(1-\pi))^{x4}(\frac{1}{4}\pi)^{x5}$$

MLE for Complete Data

$$\log L(\pi) = x1\log(\frac{1}{2})x2\log(\frac{1}{4}\pi) + x3\log(\frac{1}{4}(1-\pi)) + x4*\log(\frac{1}{4}(1-\pi))$$
$$+ x5*\log(\frac{1}{4}\pi) + \log(\frac{n!}{x1!x2!x3!x4!x5!})$$

$$\frac{d\log L(\pi)}{d\pi} = \frac{x2+x5}{\pi} - \frac{x3+x4}{1-\pi} = 0$$

$$\frac{d\log L(\pi)}{d\pi} = \frac{x2+34}{\pi} - \frac{38}{1-\pi} = 0$$

E-Step

- What are the sufficient statistics?
  - X1 (X2 can be inferred from X1, since X2 = 125-X1)

- How can their expected value be computed?
  - $E[x1|y1] = n*p(x1|y1)$

- The unobserved counts x1 and x2 are the categories with a sample size of 125
  - $p(x1) + p(x2) = p(y1) = \frac{1}{2} + \frac{1}{4} * pi$

E-Step

- $E[x1|y1] = n*p(x1|y1)$
  - $p(x1|y1) = \frac{1}{2} / (\frac{1}{2} + \frac{1}{4} * pi)$
- $E[x2|y1] = n*p(x2|y1) = 125 – E[x1|y1]$
  - $p(x2|y1) = \frac{1}{4}*pi / (\frac{1}{2} + \frac{1}{4}*pi)$
- Iteration 1? Start with pi = 0.5 (this is just a random guess)

E-Step Iteration 1

- E[x1|y1] = 125 * (½ / (½ + ¼ * 0.5)) = 100
- E[x2|y1] = 125 – 100 = 25

- These are the expected values of the sufficient statistics, given the observed data and current parameter estimates (which was just a guess)

M-Step Iteration 1

- Given sufficient statistics, make MLEs as usual

$$\frac{d \log L(\pi)}{d\pi} = \frac{x2 + 34}{\pi} - \frac{38}{1-\pi} = 0$$

$$\frac{25 + 34}{\pi} - \frac{38}{1-\pi} = 0$$

$$\pi = 0.608$$

E-Step Iteration 2

- $E[x1|y1] = 125 * (\frac{1}{2} / (\frac{1}{2} + \frac{1}{4} * 0.608)) = 95.86$
- $E[x2|y1] = 125 – 95.86 = 29.14$

- These are the expected values of the sufficient statistics, given the observed data and current parameter estimate (from iteration 1).

M-Step Iteration 2

- Given sufficient statistics, make MLEs as usual

$$\frac{d \log L(\pi)}{d\pi} = \frac{x2 + 34}{\pi} - \frac{38}{1-\pi} = 0$$

$$\frac{29.14}{\pi} - \frac{38}{1-\pi} = 0$$

$$\pi = 0.624$$

Result?

- Converge in 4 iterations to pi=0.627
  - E[x1|y1] = 95.2
  - E[x2|y1] = 29.8

Conclusion

- Distribution must be appropriate to problem
- Sufficient statistics should be identifiable and have computed expected values
- Maximization operation should be possible
- Initialization should be good or lucky to avoid saddle points and local maxima.