BackGround

Basics of Probability and Statistics

서울시립대학교 김윤나

CONTENTS

0 Intro

Ω

probability space

θ

probability distribution

$x \sim \theta$

x is drawn from theta

random variable x is drawn from the probability distribution θ

0 Intro

probability distribution

discrete probability distribution

- models only assign probabilities for a finite(discrete) set of outcomes

continuous probability distribution

- where there are an infinite number of “events” that are not countable

0 Intro

3 axioms that should be satisfied in θ with Ω

$$0 \leq p_\theta(\omega \in \Omega) \leq 1$$

Each event has a probability between zero and one

$$p_\theta(\omega') = 0, \omega' \notin \Omega \quad \text{and} \quad p_\theta(\Omega) = 1$$

event not in has Ω probability zero

the probability of any event occurring from Ω is one

$$\sum_{\omega \in \Omega} p_\theta(\omega) = 1$$

The probability of all events sums to one

1 Joint and Conditional Probabilities

Joint probability

measures the likelihood that two events occur simultaneously

$$p(x_c = green, x_s = circle) = \frac{1}{6}$$

Conditional probability

measures the likelihood that one event occurs given that another event has already occurred

$$p(X = x \mid Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)}$$

$(random\ variables\ X\ and\ Y)$

② Bayes' Rule

$$p(X \mid Y) = \frac{p(X, Y)}{p(Y)} \quad \text{and} \quad p(Y \mid X) = \frac{p(Y, X)}{p(X)}$$

$$p(X \mid Y)p(Y) = p(X, Y) = p(Y \mid X)p(X)$$

$$p(X \mid Y) = \frac{p(Y \mid X)p(X)}{p(Y)}$$

$if)\ X\ and\ Y\ are\ independent, then\ p(X|Y) = p(X)$

3 Coin Flips and the Binomial Distribution

If we want to model n throws and find the probability of k successes,

when the order of outcomes are not given

$$p(k \text{ heads}) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$$

$$p(k \text{ heads}) = \frac{n!}{k!(n-k)!}\theta^k(1-\theta)^{n-k}$$

when the order of outcomes are given

$$\theta \times (1-\theta) \times \theta \times \theta \times (1-\theta) = \theta^3(1-\theta)^2$$

$$p(k\ heads) = \theta^k(1-\theta)^{n-k}$$

4 Maximum Likelihood Parameter Estimation

Maximum Likelihood Estimation(MLE)

- to figure out what θ is based on the observed data
- choose the θ that has the highest likelihood given our data

ex) $p(D|\theta) = \theta^3(1-\theta)^2$

to find the θ that maximizes the function $f(\theta) = \theta^3(1-\theta)^2$

$$\log f(\theta) = 3log\theta + 2\log(1-\theta)$$

$$\frac{d\log f(\theta)}{d\theta} = \frac{3}{\theta} - \frac{2}{1-\theta} = 0$$

$$\theta = \frac{3}{5}$$

4 Maximum Likelihood Parameter Estimation

Generally,

$$\theta_{MLE} = \arg\max_{\theta} p(D \mid \theta)$$

$$= \arg\max_{\theta} \theta^{H}(1-\theta)^{T}$$

$$= \frac{H}{H+T}.$$

The value of an arg max expression stays the same
if we perform any monotonic transformation of the function
inside arg max

5 Bayesian Parameter Estimation

Potential problem of MLE

→ often inaccurate when the size of the data sample is small since it always attempts to fit the data as well as possible

Problem of *"overfitting"*

This can be addressed and alleviated by considering the *uncertainty* on the parameter and using Bayesian parameter estimation instead of MLE

In Bayesian parameter estimation, we consider a distribution over all the possible values for the parameter.

5 Bayesian Parameter Estimation

$p(\theta)$: to represent a distribution over all possible values for θ

which encodes our prior belief about what value is the true value of θ

D : data D provide evidence for or against that belief

$$p(\theta \mid D) = \frac{p(D \mid \theta)p(\theta)}{p(D)} \qquad (by\ Bayes'\ rule)$$

$$p(D) = \int_{\theta'} p(\theta')p(D \mid \theta')d\theta' \qquad (for\ continuous\ distribution)$$

$$p(\theta \mid D) = \frac{p(D \mid \theta)p(\theta)}{\int_{\theta'} p(\theta')p(D \mid \theta')d\theta'} \qquad the\ probability\ for\ a\ particular\ \theta$$

5 Bayesian Parameter Estimation

$$p(\theta \mid D) = \frac{p(D \mid \theta)p(\theta)}{\int_{\theta'} p(\theta')p(D \mid \theta')d\theta'}$$

- $p(\theta \mid D)$: the posterior probability of θ
- $p(\theta)$: the prior probability of θ
- $p(D \mid \theta)$: the likelihood of D
- $\int_{\theta'} p(\theta')p(D \mid \theta')d\theta'$: the marginal likelihood of D

$$p(D) = \int_{\theta'} p(\theta')p(D \mid \theta')d\theta' \qquad (for\ continuous\ distribution)$$

5 Bayesian Parameter Estimation

$p(\theta \mid D) \propto p(\theta)p(D \mid \theta)$ since the likelihood of the data remains constant

→ proportional to the prior times the likelihood

What is a Posterior Distribution?

The posterior distribution is a way to summarize what we know about uncertain quantities in Bayesian analysis. It is a combination of the prior distribution and the likelihood function, which tells you what information is contained in your observed data (the "new evidence"). In other words, **the posterior distribution summarizes what you know after the data has been observed**. The summary of the evidence from the new observations is the likelihood function.

Posterior Distribution = Prior Distribution + Likelihood Function ("new evidence")

Posterior distributions are vitally important in Bayesian Analysis. They are in many ways the goal of the analysis and can give you:

- Interval estimates for parameters,
- Point estimates for parameters,
- Prediction inference for future data,
- Probabilistic evaluations for your hypothesis.

5 Bayesian Parameter Estimation

To compute the mean of the posterior distribution, which is given by the weighted sum of probabilities and the parameter values.

$$E[X] = \sum_x x p(x) \qquad for\ discrete\ distribution$$

$$E[X] = \int_x x f(x) dx \qquad for\ continuous\ distribution$$

$$\theta_{MAP} = \arg\max_\theta p(\theta \mid D) = \arg\max_\theta p(D \mid \theta) p(\theta)$$

Sometimes, we are interested in using the mode of the posterior distribution as our estimate of the parameter, called Maximum a Posteriori(MAP)

6 Probabilistic Models and Their Applications

In the case of a distribution over words, we have parameter for each element in V

<Workflow>

1. Define the model
   - Capture the probabilities
2. Learn its parameters
   - Figure out actually how to set the probabilities for each word
3. Apply the model
   - Once θ is defined, analyze the probability of a specific subset of words in the corpus and observe unseen data & calculating the probability of seeing the words in the new text.