Probability theory

LING 570

Fei Xia

Week 1: 09/30/09

Basic concepts

- Possible outcomes, sample space, event, event space

- Random variable and random vector

- Conditional probability, joint probability, marginal probability (prior)

Random variable

- The outcome of an experiment need not be a number.
- We often want to represent outcomes as numbers.
- A random variable X is a function: $\Omega \rightarrow R$.
  - Ex: the number of heads with three tosses: X(HHT)=2, X(HTH)=2, X(HTT)=1, …

Two types of random variables

- Discrete: X takes on only a countable number of possible values.
  - Ex: Toss a coin three times. X is the number of heads that are noted.

- Continuous: X takes on an uncountable number of possible values.
  - Ex: X is the speed of a car (e.g., 56.5 mph)

Common distributions

- Discrete random variables:
  - Uniform
  - Bernoulli
  - binomial
  - multinomial
  - Poisson

- Continuous random variables:
  - Uniform
  - Gaussian

Random vector

- Random vector is a finite-dimensional vector of random variables: $X=[X_1,\ldots,X_k]$.

- $P(x) = P(x_1,x_2,\ldots,x_n)=P(X_1=x_1,\ldots., X_n=x_n)$

- Ex: $P(w_1, \ldots, w_n, t_1, \ldots, t_n)$

Notation

- X, Y: random variables or random vectors.
- x, y: some values

- P(X=x) is often written as P(x)
- P(X=x | Y=y) is written as P(x | y)

Three types of probability

- Joint prob $P(x,y)$: the prob of $X=x$ and $Y=y$ happening together
- Conditional prob $P(x \mid y)$: the prob of $X=x$ given a specific value of $Y=y$
- Marginal prob $P(x)$: the prob of $X=x$ for all possible values of Y.

Chain rule: calc joint prob from marginal and conditional prob

$$P(A,B) = P(A) * P(B \mid A) = P(B) * P(A \mid B)$$

$$P(A_1,\ldots,A_n) = \prod_{i>=1} P(A_i \mid A_1,\ldots A_{i-1})$$

Calc marginal prob from joint prob

$$P(A) = \sum_{B} P(A, B)$$

$$P(A_1) = \sum_{A_2, \ldots, A_n} P(A_1, \ldots, A_n)$$

Bayes' rule

$$P(B \mid A) = \frac{P(A,B)}{P(A)} = \frac{P(A \mid B)P(B)}{P(A)}$$

$$\begin{aligned} y^* &= \arg\max_y P(y \mid x) \\ &= \arg\max_y \frac{P(x \mid y)P(y)}{P(x)} \\ &= \arg\max_y P(x \mid y)P(y) \end{aligned}$$

Independent random variables

- Two random variables X and Y are independent iff the value of X has no influence on the value of Y and vice versa.
- $P(X,Y) = P(X)\,P(Y)$
- $P(Y|X) = P(Y)$
- $P(X|Y) = P(X)$

Conditional independence

Once we know C, the value of A does not affect the value of B and vice versa.

- $P(A,B \mid C) = P(A|C)\, P(B|C)$
- $P(A|B,C) = P(A \mid C)$
- $P(B|A, C) = P(B \mid C)$

Independence and conditional independence

- If A and B are independent, are they conditional independent?

- Example:
  - Burglar, Earthquake
  - Alarm

Independence assumption

$$P(A_1,\ldots,A_n) = \prod_{i>=1} P(A_i \mid A_1,\ldots A_{i-1})$$

$$\approx \prod_{i>=1} P(A_i \mid A_{i-1})$$

An example

- $P(w_1\ w_2\ \ldots\ w_n)$
  $= P(w_1)\ P(w_2 \mid w_1)\ P(w_3 \mid w_1\ w_2) * \ldots$
  $* P(w_n \mid w_1\ \ldots,\ w_{n-1})$
  $\approx P(w_1)\ P(w_2 \mid w_1)\ \ldots.\ P(w_n \mid w_{n-1})$

- Why do we make independence assumption which we know are not true?

Summary of elementary probability theory

- Basic concepts: sample space, event space, random variable, random vector

- Joint / conditional /marginal probability

- Independence and conditional independence

- Four common tricks:
  - Chain rule
  - Calculating marginal probability from joint probability
  - Bayes' rule
  - Independence assumption

Additional slides

Sample space, event, event space

- Sample space (Ω): the set of all possible outcomes.
  - Ex: toss a coin three times:
    {HHH, HHT, HTH, HTT, …}
- Event: an event is a subset of Ω.
  - Ex: an event is {HHT, HTH, THH}
- Event space (2^{Ω}): the set of all possible events.

Probability function

- A probability function (a.k.a. a probability distribution) distributes a probability mass of 1 throughout the sample space Ω.

- It is a function from $2^{\Omega} \rightarrow [0,1]$ such that:

  $P(\Omega) = 1$

  For any disjoint sets $A_j \in 2^{\Omega}$, $P(\bigcup A_j) = \sum P(A_j)$

  - Ex: P({HHT, HTH, HTT})

    = P({HHT}) + P({HTH}) + P({HTT})

The coin example

- The prob of getting a head is 0.1 for one toss. What is the prob of getting two heads out of three tosses?

- P("Getting two heads")

  = P({HHT, HTH, THH})

  = P(HHT) + P(HTH) + P(THH)

  = 0.1*0.1*0.9 + 0.1*0.9*0.1+0.9*0.1*0.1

  = 3*0.1*0.1*0.9

The coin example (cont)

- X = the number of heads with three tosses

- P(X=2)

  $= P(\{HHT, HTH, THH\})$

  $= P(\{HHT\}) + P(\{HTH\}) + P(\{THH\})$

Maximum likelihood estimation

- An example: toss a coin 3 times, and got two heads. What is the probability of getting a head with one toss?

- Maximum likelihood: (ML)

  $\theta^* = \arg\max_\theta P(\text{data} \mid \theta)$

- In the example,
  - P(X=2) = 3 * p * p * (1-p)

    e.g., the prob is 3/8 when p=1/2, and is 12/27 when p=2/3

    3/8 < 12/27

    ➔ when p=2/3, P(X=2) reaches the maximum.