UNIVERSITY of DELAWARE

Basics of Probability

(Search and Data Mining)

Hui Fang
Department of Electrical and Computer Engineering
University of Delaware

UNIVERSITY of DELAWARE

Why probability?

- Probability provides a principled way to quantify the uncertainties associated with natural language
- Example applications
  - Given that we observe "baseball" three times in a news article, how likely is it about "sports"? *Text categorization*
  - Given that a user is interested in sports news, how likely would the user use "baseball" in a query? *Search*

UNIVERSITY of DELAWARE

Basic Concepts in Probability

- Experiment
  - A procedure that yields one of a given set of possible outcomes
- Sample space
  - The set of possible outcomes for the experiment
- Event
  - A subset of sample space
  - We use a Random Variable to denote an event, and it is often a capital letter, such as A.
- Probability P(A)
  - The fraction of possible worlds in which A is true

UNIVERSITY of DELAWARE

An Example

- Experiment
  - Roll a single 6-sided die one time
- Sample space
  - {1,2,3,4,5,6}
- Random Variable/Event
  - A = Roll an even number {2,4,6}
- Probability P(A)
  - $\frac{|\{2,4,6\}|}{|\{1,2,3,4,5,6\}|} = \frac{3}{6}$

UNIVERSITY of DELAWARE

Probability Distributions

- A probability is a single number.

  $P(W = rainy) = 0.1 \qquad P(rainy) = 0.1$
- A distribution is a TABLE of probabilities of values

W	P(W)
sunny	0.6
rainy	0.1
cloudy	0.3

- Properties:

  $\forall x, P(x) \geq 0 \qquad \sum_x P(x) = 1$

UNIVERSITY of DELAWARE

Joint Distributions

- A joint distribution measures the likelihood that multiple events occur simultaneously.

T	W	P(T,W)
hot	sunny	0.4
hot	rainy	0.1
cold	sunny	0.2
cold	rainy	0.3

P(T=hot, W=rainy) = 0.1

- Notation: $P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = P(x_1, x_2, ..., x_n)$
- Properties:

  $P(x_1, x_2, ..., x_n) \geq 0, \quad \sum_{(x_1, x_2, ..., x_n)} P(x_1, x_2, ..., x_n) = 1$

Marginal Distributions

- *Marginal distributions* are sub-tables which eliminate variables through marginalization (e.g., summing out).
  - Combine collapsed rows by adding

T	W	P(T,W)
hot	sunny	0.4
hot	rainy	0.1
cold	sunny	0.2
cold	rainy	0.3

T	P(T)
hot	0.5
cold	0.5

$$P(X_1 = x_1) = \sum_{x_2} P(X_1 = x_1, X_2 = x_2)$$

Conditional Probabilities

- A conditional probability measures the likelihood that one event occurs given that another event has already occurred.

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

Conditional Probabilities - Example

Joint probabilities

T	W	P(T,W)
hot	sunny	0.4
hot	rainy	0.1
cold	sunny	0.2
cold	rainy	0.3

Conditional probability

P(T=hot | W = rainy) = ???

$$P(T = hot \mid W = rainy) = \frac{P(T = hot, W = rainy)}{P(W = rainy)}$$

$$= \frac{P(T = hot, W = rainy)}{P(T = hot, W = rainy) + P(T = cold, W = rainy)}$$

$$= \frac{0.1}{0.1 + 0.3}$$

Probabilistic Inference

- It is to compute a desired probability from other known probabilities
- We generally compute conditional probabilities
  - Given an evidence, compute the probability of a belief
  - For example,
    P(on time | no accidents) = 0.90
- Probabilities change with new evidence:
  - P(on time | no accidents, 5 a.m.) = 0.95
  - P(on time | no accidents, 5 a.m., raining) = 0.80

Independence

- Two variables are independent in a joint distribution if:

$$P(X,Y) = P(X)P(Y)$$

- Independence is often used as a simplifying assumption.

Bayes' Rule

- Two ways to factor a joint distribution over two variables:

$$P(X \mid Y) = \frac{P(X,Y)}{P(Y)} \qquad P(Y \mid X) = \frac{P(X,Y)}{P(X)}$$

- Rewriting, we get:

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

- Dividing, we get:

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

Bayes' Rule - Example

If you wake up with a headache, how likely do you have flu?

- H denotes "having headache"
- F denotes "having flu".
- The problem is to compute P(F|H).

Assume we know

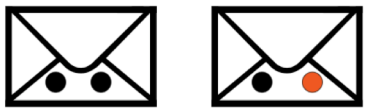
- *One in ten people has headache*
- *One in 100 people has flu*
- *90% of people who have flu have headache*

$$P(F|H) = \frac{P(H|F) \times P(F)}{P(H)}$$

$$= \frac{0.9 * 0.01}{0.1} = 0.09$$

Another Example

- There are two identical-looking envelopes
  - One has a red ball (worth $100) and a black ball
  - One has two black balls. Black balls worth nothing



- You randomly grabbed an envelope, randomly took out one ball – it is black.
- As this point, you are given the option to switch the envelope. To switch or not to switch?

Inference with Bayes' Rule – An Example

- A given coin is either fair or has a 60% bias in favor of head. We know that the probability of getting a fair coin is 0.75. Now, we have tossed a coin and get a head, how likely is the coin the fair one?
- Formally, the problem can be summarized as :
  - There are two hypotheses h1 and h2
    - h1: a fair coin : P(Head|h1)=0.5;
    - h2: a biased coin: P(Head|h2)=0.6.
  - Prior probabilities: P(h1)=0.75, P(h2)=0.25.
  - We want to compute P(h1|Head).
- Solution:
  - Using Bayes' rule, we get P(h1|Head) = P(Head|h1)* P(h1)/P(Head).
  - Since P(Head)=P(Head|h1)*P(h1)+P(Head|h2)*P(h2)=0.525, so we have P(h1|Head) = 0.714

Inference with Bayes' Rule

Hypothesis space: H={H1 , …, Hn} Evidence: E

$$P(H_i \mid E) = \frac{P(E \mid H_i)P(H_i)}{P(E)}$$

If we want to pick the most likely hypothesis H*, we can drop P(E)

Posterior probability of Hi — Prior probability of Hi

$$P(H_i \mid E) \propto P(E \mid H_i)P(H_i)$$

Likelihood of data/evidence if Hi is true

Model for Coin Flips

- Two possible outcomes in coin flipping: h (head) or t (tail).
- Notations: $P(h) = \theta$, so $P(t) = 1 - \theta$.

- Assuming all the flips are independent, what is the probability of observing the particular sequence of outcomes (h,t,h,h,t)?

$$P(h,t,h,h,t) = P(h) \times P(t) \times P(h) \times P(h) \times P(t)$$
$$= \theta \times (1-\theta) \times \theta \times \theta \times (1-\theta)$$
$$= \theta^3 \times (1-\theta)^2$$

Parameter Estimation for Coin Flipping

- Given we observe the data D, where D={h,t,h,h,t}, we want to figure out the value of θ .
- **Maximum Likelihood Estimation:**
  - **Choose the parameter value that has the highest likelihood given our data.**
- The probability of the data observed is: $P(D|\theta) = \theta^3 \times (1-\theta)^2$
- So, we want to find the parameter value that maximizes this quantity.

$$\theta_{MLE} = \arg\max_{\theta} P(D|\theta) = \arg\max_{\theta} \log(P(D|\theta)) = \arg\max_{\theta} f(\theta)$$

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

Parameter Estimation for Coin Flipping (Cont.)

- So, we want to find the parameter value that maximizes this quantity.

$$\theta_{MLE} = \arg\max_{\theta} P(D|\theta) = \arg\max_{\theta} \log(P(D|\theta)) = \arg\max_{\theta} f(\theta)$$

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

- Taking derivative and set the value to 0, we get

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

- So, the solution is $\theta = \frac{3}{5}$
- More generally, for H heads, T tails, we have $\theta = \frac{H}{H+T}$

Statistical Learning

- Learning a model means estimating its parameters.
- General workflow:
  - Define a model
  - Learn its parameters
  - Apply the model