Probabilistic Modeling

and

Joint Distribution Model

Elements of Probability Theory

Introduction

- Concerned with analysis of random phenomena
- Originated from gambling & games
- Uses ideas of counting, combinatorics and measure theory
- Uses mathematical abstractions of non-deterministic events

Elements of Probability Theory

Introduction

- Continuous probability theory deals with events that occur in a continuous sample space
- Discrete probability deals with events that occur in countable sample spaces
- Events : a set of outcomes of an experiment
- Events : a subset of sample space

Elements of Probability Theory

Axioms of Probability

- **Nonnegativity :** $0 \leq P(E) \leq 1$
- **Additivity :** $P(E_1, E_2 \ldots, E_n) = \sum_i P(E_i)$
- **Normalization (unit measure):** $P(\Omega) = 1$, $P(\varnothing) = 0$

- **Some consequences:**
- $P(\Omega \setminus E) = 1 - P(E)$ $\{\Omega$: universe$\}$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A \setminus B) = P(A) - P(B)$ if $B \subseteq A$

Elements of Probability Theory

Conditional probability

- **Bayes Rule :** P(A|B) =P(A,B) / P(B)
- **OR :**
- P(A|B) =P(B|A).P(A) / P(B)
- **Independency condition :** P(A,B) = P(A).P(B)
- **Mutually exclusive events :** P(A,B) = 0
- **Mutually exclusive events :** P(A U B) = P(A) + P(B)
- OR
- P(A \ B) = P(A)

Elements of Probability Theory

Random Variables

- A variable
- A function mapping the sample space of a random process to the values
- Values can be discrete or continuous
- Each outcome as value (or a range) is assigned a probability

Elements of Probability Theory

Random Variables

- A variable
- A function mapping the sample space of a random process to the values
- Values can be discrete or continuous
- Discrete example : fair coin toss
- X={ 1 if heads, 0 if tails}
- Or fair dice roll : X={ "the number shown on dice"}

Elements of Probability Theory

Random Variables

- Continuous example: spinner
- Outcome can be any real number in $[0,2\pi)$
- Any specific value has zero probability
- So we use ranges instead of single points
- E.g. having a value in $[0,\pi/2]$ has probability 1/4

Elements of Probability Theory

Random Variables

- In case of discrete random variables we use probability mass function
- $P_X(x) = \left\{ \text{1/2 if X=0, 1/2 if X=1, 0 otherwise} \right\}$

- Notice the use of uppercase for the random variable and lowercase for the mass function variable
- Cumulative distribution function (CDF) :

$$F_X(x) = P(X \leq x)$$

Elements of Probability Theory

Random Variables

- In case of continuous variables,
- We use a probability density function

$$P_X[a \le X \le b] = \int_a^b p(x)dx$$

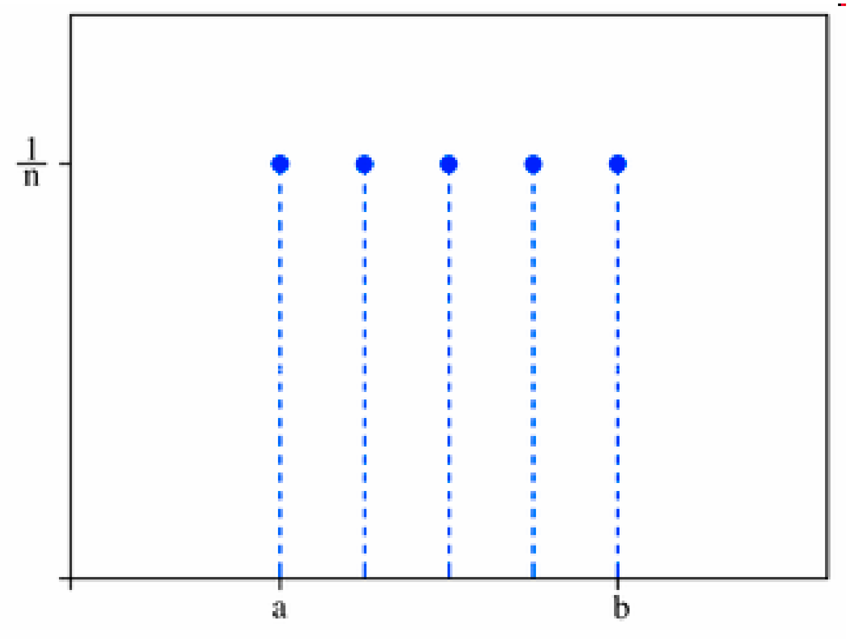
- So that the CDF becomes

$$F_X(x) = \int_{-\infty}^{x} p(u)du$$

Elements of Probability Theory

Well Known Distributions

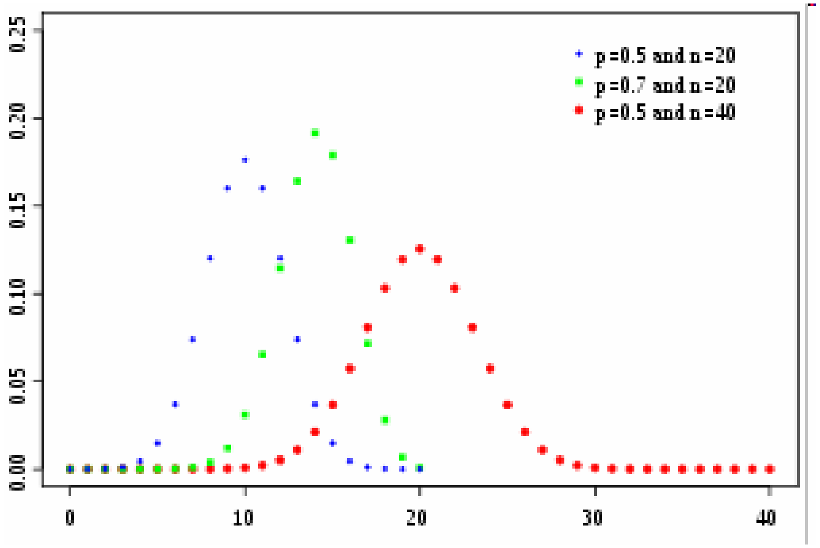
- Discrete uniform distribution



Elements of Probability Theory

Well Known Distributions

- Binomial distribution $$\Pr(K=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

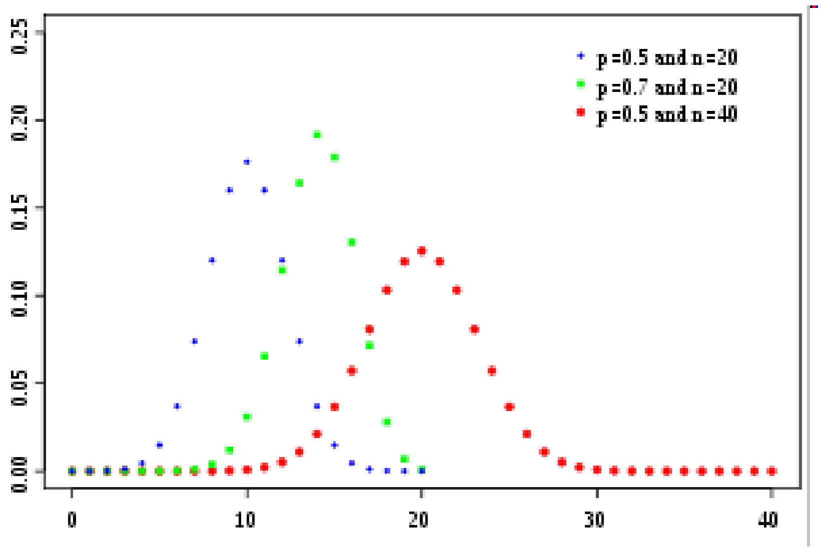


- Special case : n=1 -> Bernoulli distribution

Elements of Probability Theory

Well Known Distributions

- Special case : n=1 -> Bernoulli distribution

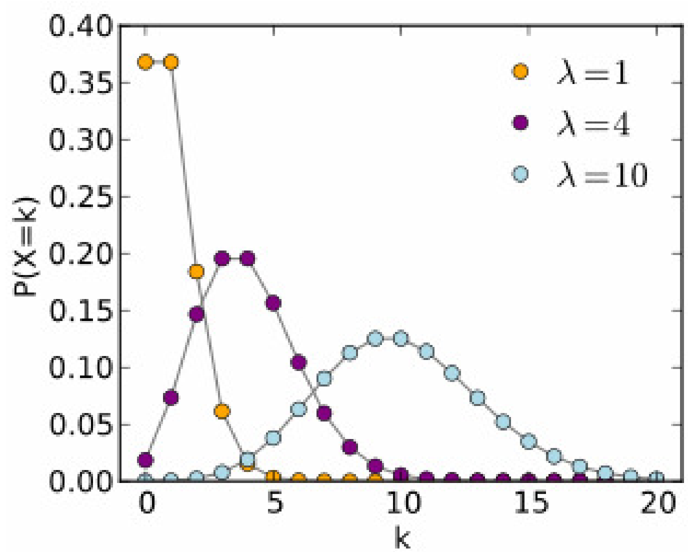


Elements of Probability Theory

Well Known Distributions

- Poisson distribution : n events occur with a known average rate λ and independently of the time since the last event

$$f(n;\lambda) = \frac{\lambda^n e^{-\lambda}}{n!}$$



Elements of Probability Theory

Expected Value and Variance

- Expected value : A measure of probability weighted average of expected outcomes

$$\mathrm{E}(X) = \sum_i x_i p(x_i) \qquad \mathrm{E}(X) = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x$$

- Variance : expected value of the square of the deviation of random variable from its expected value

$$\mathrm{Var}(X) = \mathrm{E}[(X-\mu)^2] \quad \mathrm{Var}(X) = \int (x-\mu)^2 p(x)\,dx$$

Elements of Probability Theory

Joint Distributions

- More than one random variable
- On the same probability space (universe)
- Events defined in terms of all variables
- Called multivariate distribution
- Called bivariate if two variables involved
- Remembering Bayes rule, conditional distribution:

$$\begin{aligned} \mathrm{P}(X = x \text{ and } Y = y) &= \mathrm{P}(Y = y \mid X = x) \cdot \mathrm{P}(X = x) \\ &= \mathrm{P}(X = x \mid Y = y) \cdot \mathrm{P}(Y = y). \end{aligned}$$

Probabilistic Modeling

Joint Distributions

- Similar to probabilities, if variables are independent:

$$P(X = x \text{ and } Y = y) = P(X = x) \cdot P(Y = y)$$

- Continuous distribution case:

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$$

- Marginal distributions:

$$\mathrm{P}(X = x) = \sum_y \mathrm{P}(X = x, Y = y) = \sum_y \mathrm{P}(X = x|Y = y) \ \mathrm{P}(Y = y)$$

$$p_X(x) = \int_y p_{X,Y}(x, y)\, dy = \int_y p_{X|Y}(x|y)\, p_Y(y)\, dy$$

- Reduces to simple product summation if independent

Probabilistic Modeling

Random Configurations

- In general a set of n random variables:

$$V = (V_1, V_2 ..., V_n)$$

- With possible outcomes for each variable:

$$\{x_1, x_2 ..., x_m\}$$

- A configuration is a vector of x where each value is assigned to a variable

$$x = (x_1, x_2 ..., x_n)$$

CSCI 6509 Notes Fall 2009

Faculty of Computer Science Dalhousie University

Probabilistic Modeling

Random Configurations

- In modeling we assume a sequence of configurations:

$$x^{(1)}, \ldots, x^{(t)}$$

$$x^{(1)} = (x_{11}, x_{12}, \ldots x_{1n})$$

$$x^{(2)} = (x_{21}, x_{22}, \ldots x_{2n})$$

$$x^{(t)} = (x_{t1}, x_{t2}, \ldots x_{tn})$$

- Here we assume a fixed number (n) of components in each configuration, and x_{ij} are values from finite set $\{x_1, x_2 \ldots, x_m\}$

Probabilistic Modeling

Random Configurations

- NLP uses probabilistic modeling as a framework for solving problems
- Computational tasks:
  - Representation of models
  - Simulation : generating random configurations
  - Evaluation : computing probability of a complete configuration
  - Marginalization : computing probability of a partial configuration
  - Conditioning : computing conditional probability of completion given partial observation
  - Completion : find most probable completion of partial observation
  - Learning : parameter estimation

Probabilistic Modeling

Joint distribution model

- A joint probability distribution $P(X_1 = x_1, X_2 = x_2, \ldots X_n = x_n)$ specifies the probability of each complete configuration $\mathbf{x} = (x_1, x_2 \ldots, x_n)$

- In general it takes m x n parameters (less one constraint) to specify an arbitrary joint distribution on n random variables with m values

Probabilistic Modeling

Joint distribution model

- This can be captured in lookup table $\theta_{x^{(1)}}, \theta_{x^{(1)}}, \ldots \theta_{x^{(V^n)}}$ where $\theta_{x^{(k)}}$ gives the probability of RV's taking on jointly the configuration $x^{(k)}$
- So $\theta_{x^{(k)}} = P(X = x^{(k)})$
- Satisfying $\sum_{k=1}^{V} \theta_{x^{(k)}} = 1$

Probabilistic Modeling

More on computational tasks

- **Simulation** : Given the lookup table representation, compute the cumulative value $\theta_{x^{(k)}}$ of the configurations, select the $x^{(k)}$ whose cumulative probability interval contains a given p value
- **Evaluation** : Evaluate the probability of a complete configuration

  $$x = (x_1, x_2 \ldots, x_n)$$

  From the lookup table: $P(X_1 = x_1, \ldots X_n = x_n) = \theta_{(x_1 x_2 \ldots\ldots x_n)}$
- **Marginalization** : the probability of an incomplete configuration:

  $$P(X_1 = x_1, \ldots X_n = x_n) = \sum_{y_{k+1}} \ldots \sum_{y_n} P(X_1 = x_1, \ldots X_k = x_k, X_{k+1} = y_{k+1} \ldots, X_n = y_n)$$

  From lookup table: $= \sum_{y_{k+1}} \ldots \sum_{y_n} \theta_{(x_1 x_2 \ldots\ldots x_k, y_{k+1} \ldots\ldots y_n)}$

Probabilistic Modeling

More on computational tasks

- **Completion :** Compute the conditional probability of a possible completion $(y_{k+1}, y_{k+2} \dots, y_n)$
  given an incomplete configuration $x = (x_1, x_2 \dots, x_n)$

  Need to evaluate a complete configuration and then divide by a marginal sum

$$\frac{\theta_{(x_1 x_2 \dots\dots x_k\, y_{k+1} \dots\dots y_n)}}{\sum_{z_{k+1}} \dots \sum_{z_n} \theta_{(x_1 x_2 \dots\dots x_k\, , z_{k+1} \dots\dots z_n)}}$$

Probabilistic Modeling

Example

- Spam detection : an arbitrary e-mail message is spam or not

- Caps = ‘Y’ if the message subject line does not contain lowercase letter, ‘N’ otherwise,
- Free = ‘Y’ if the word ‘free’ appears in the message subject line (letter case is ignored), ‘N’ otherwise,

  and
- Spam = ‘Y’ if the message is spam, and ‘N’ otherwise.

Randomly select 100 messages, count how many times each configuration appears

Probabilistic Modeling

Example

- Given a fully specified joint distribution table, one can lookup the probability of any configuration. For example:

  P(Free = Y; Caps = Y; Spam = Y) = 0.2

  P(Free = Y; Caps = N; Spam = N) = 0.0

Free	*Caps*	*Spam*	Number of messages	Estimated probability
Y	Y	Y	20	0.20
Y	Y	N	1	0.01
Y	N	Y	5	0.05
Y	N	N	0	0.00
N	Y	Y	20	0.20
N	Y	N	3	0.03
N	N	Y	2	0.02
N	N	N	49	0.49
		Total:	100	1.0

Probabilistic Modeling

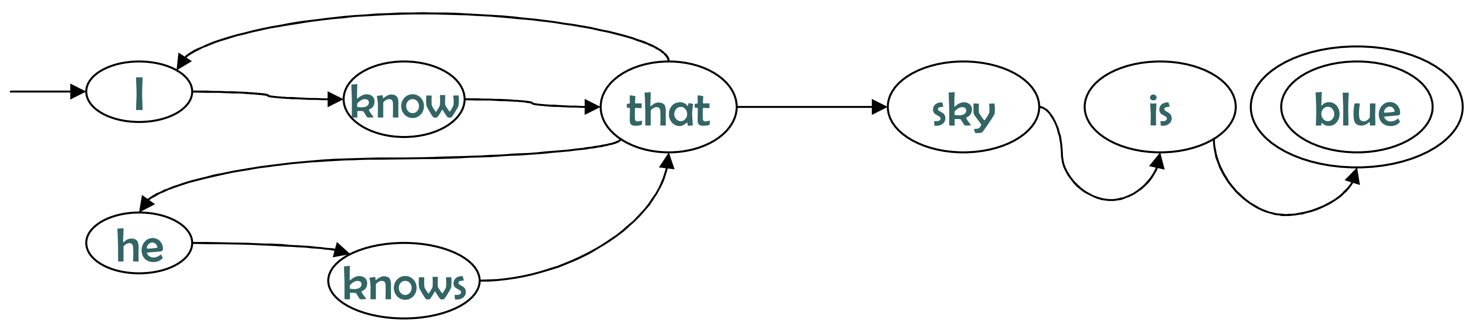
Joint distribution model

- Drawbacks of Joint Distribution Model:
  - memory cost to store table
  - running-time cost to do summations
  - the sparse data problem in learning

Probabilistic Modeling

Generative Model

- Idea for traditional generative model:
- what does the automaton below <u>generate</u> ?

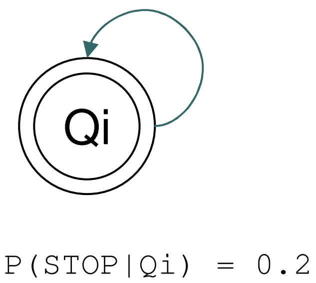


- I know that sky is blue, I know that he knows that sky is blue, I know that I know that sky is blue, …
- But not : sky is blue, I know he, I blue that…
- This is the language of this automaton

Probabilistic Modeling

Generative Model

- Idea for probabilistic generative model:



(Manning, Raghavan & Schutze, 2009)

string	assigned probability
the	0.2
a	0.1
frog	0.01
toad	0.01
said	0.03
likes	0.02
that	0.04
....	....

- If instead each node has a probability distribution over generating different terms, we have a language model

Probabilistic Modeling

Generative Model

- A language model is a function that puts a probability measure over strings drawn from some vocabulary

$$\sum_{i \in \Sigma^*} P(t_i) = 1$$

- Each $P(t_i)$ is a term emission probability in this unigram model
- Such a model places a probability distribution over any sequence of words
- By construction, it also provides a model for generating text according to its distribution

Probabilistic Modeling

Generative Model

- P(frog said that toad likes frog) = (0.01 ×0.03 × 0.04 × 0.01 × 0.02 × 0.01) {emission probabilities}

  X (0.8 ×0.8 × 0.8 × 0.8 × 0.8 × 0.8 × 0.2) {continue/stop probabilities}

  = 0.000000000001573
- Usually continue/stop probabilities are omitted when comparing models
- Based on computed value, a model is more likely

Probabilistic Modeling

Generative Model

- Compare this model to the previous model:

string	assigned probability
the	0.15
a	0.12
frog	0.0002
toad	0.0001
said	0.03
likes	0.04
that	0.04
....	....

{omitting P(stop)}

P(s|M1) = 0.0000000000048

P(s|M2) = 0.000000000000000384

So model 1 is more likely

Probabilistic Modeling

Types of Generative Models

- In general for a sequence of events using earlier successive events using *Bayesian Inference Rule*:

$$P(t_1 t_2 t_3 t_4) = P(t_1) P(t_2 | t_1) P(t_3 | t_1 t_2) P(t_4 | t_1 t_2 t_3)$$

- If total independence among events exists:

$$P_{uni}(t_1 t_2 t_3 t_4) = P(t_1) P(t_2) P(t_3) P(t_4)$$

- This is unigram model

Probabilistic Modeling

Types of Generative Models

- If only conditioning is on the previous term

$$P_{bi}(t_1 t_2 t_3 t_4) = P(t_1)P(t_2 \mid t_1)P(t_3 \mid t_2)P(t_4 \mid t_3)$$

- This is bigram model

- Unigram models frequently used when sentence structure is not important
- E.g. in IR but not in speech recognition

Probabilistic Modeling

Types of Generative Models

- Unigram models are of type 'bag of words'

$$P_{bi}(t_1 t_2 t_3 t_4) = P(t_1)P(t_2 \mid t_1)P(t_3 \mid t_2)P(t_4 \mid t_3)$$

- Recalls a multinomial distribution of probabilities over words

$$P(d) = \frac{L_d\,!}{tf_{t_1,d}\,!\,tf_{t_2,d}\,!\ldots tf_{t_M,d}\,!} P(t_1)^{f_{t1,d}}\, P(t_2)^{f_{t2,d}} \ldots. P(t_M)^{f_{tM,d}}$$

- Where L_d is the length of document d with vocabulary of size M
- Observe here the positions of the terms are insignificant

Probabilistic Modeling

Types of Generative Models

- Fundamental question: which model to use?
- Speech recognition: the model has to be general enough beyond observed data to allow unknown sequences
- IR : a document is finite and mostly fixed
  - Get a representative sample
  - Build a language model for document
  - Calculate generative probabilities of sequences from the model
  - Rank documents by probability ranking principle

Probabilistic Approaches

Probability Ranking Principle

- rank documents by their estimated probability of relevance
- $P(R = 1|d, q)$ for document d, query q
- Basic case : 1/0 loss
- Rank documents, return top k
- Non restrictive case : Bayes optimal decision rule
- d is relevant iff $P(R = 1|d, q) > P(R = 0|d, q)$

Probabilistic Approaches

Probability Ranking Principle

- If cost is involved:

$$C0 \cdot P(R = 0|d) - C1 \cdot P(R = 1|d)$$
$$\leq$$
$$C0 \cdot P(R = 0|d') - C1 \cdot P(R = 1|d'))$$

where

C1 = cost of missing relevant document

C0 = cost of returning nonrelevant document

Probabilistic Modeling

Types of Other Generative Models

- Rather than a document model, and checking likelihood of generating query,
- Build a query model and check likelihood of generating a document
- OR: use both approaches together
  - Needs a measure of divergence between document and query models
  - Kullback-Leibler divergence:

$$R(d;q) = \sum_{t \in V} P(t \mid M_q) \log \frac{P(t \mid M_q)}{P(t \mid M_d)}$$

Probabilistic Modeling

Types of Other Generative Models

- Translational model generates query words not in a document by translating into alternate terms with similar meaning,
- Needs to know conditional probability distribution between vocabulary terms

$$P(q \mid M_d) = \prod_{t \in q} \sum_{v \in V} P(v \mid M_d) T(t \mid v)$$

  - Where $P(q \mid M_d)$ is the query translation model, $P(v \mid M_d)$ is the document language model, $T(t \mid v)$ is the conditional probability distribution between vocabulary terms

Sources :

CSCI 6509 Notes Fall 2009

Faculty of Computer Science Dalhousie University
http://www.cs.dal.ca/~vlado/csci6509/coursecalendar.html

Manning, Raghavan & Schutze, 2009, An introduction to information retrieval

Jurafsky, Martin, 2000, An Introduction to NLP,Computational Linguistics and Speech Recognition

Ghahramani,2000, Fundamentals of Probability