UVA CS 6316 – Fall 2015 Graduate: Machine Learning

Lecture 11: Probability Review

Dr. Yanjun Qi

University of Virginia

Department of Computer Science

Announcements: Schedule

- Midterm – Nov. 18th or 19th / 3:30pm – 4:45pm / open notes
- HW4 is totally for sample midterm questions
- HW4 will be out next Wed, due on Nov 16th (i.e. for a good preparation for midterm. Solution will be released before due time.)

- HW3 will be out next Monday, due on Nov 6th
- Project proposal is due on Oct 11 midnight (2 pages min.)

- Grading of HW1 is available on Collab
- Solution of HW1 is available on Collab
- Grading of HW2 will be available next week
- Solution of HW2 will be available next week

Announcements: Proposal

- The following sections are expected in the proposal ,
  - Abstract / Summary (300 words limit)
  - Introduction of the target task
  - Previous solutions for the target task
  - Why is this related to machine learning ?
  - Proposed Method (optional for this phase)
  - Experimental design / where to get data / data statistics / details of the data ?
  - Why are you the right person/team for implementing this plan ?
  - References (not included in the page counts)

Please use the provided template (in collab proposal page);

Where are we ? ➔ Five major sections of this course

- ❑ Regression (supervised)
- ❑ Classification (supervised)
- ❑ Unsupervised models
- ❑ Learning theory
- ❑ Graphical models

Where are we ? ➔

Three major sections for classification

- We can divide the large variety of classification approaches into roughly three major types

1. Discriminative
   - directly estimate a decision rule/boundary
   - e.g., support vector machine, decision tree

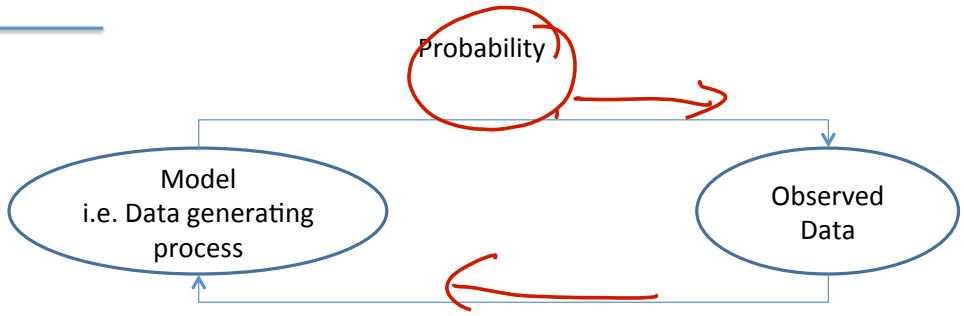
2. Generative:
   - build a generative statistical model
   - e.g., naïve bayes classifier, Bayesian networks

3. Instance based classifiers
   - Use observation directly (no models)
   - e.g. K nearest neighbors

Today : Probability Review

- The big picture
- Events and Event spaces
- Random variables
- Joint probability, Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
- Structural properties
  - Independence, conditional independence

The Big Picture



Estimation / learning / Inference / Data mining

But how to specify a model?

Probability as frequency

- Consider the following questions:
  - 1. What is the probability that ~~when I flip~~ a coin it is "heads"?
  - 2. why ?

We can count ➔ ~1/2

  - 3. What is the probability of Blue Ridge Mountains to have an erupting volcano in the near future ?

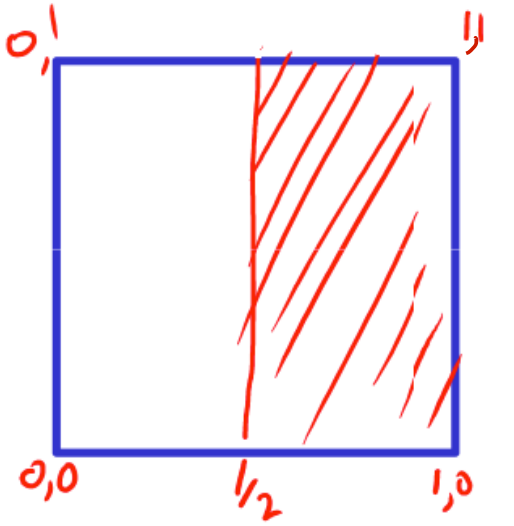
➔ could not count

Message: *The frequentist view is very useful, but it seems that we also use domain knowledge to come up with probabilities.*

Adapt from Prof. Nando de Freitas's review slides

Probability as a measure of uncertainty

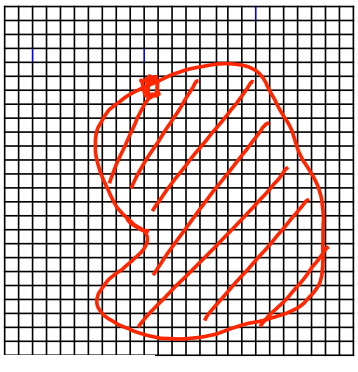
- Imagine we are throwing darts at a wall of size 1x1 and that all darts are guaranteed to fall within this 1x1 wall.
- What is the probability that a dart will hit the shaded area?



Adapt from Prof. Nando de Freitas's review slides

Probability as a measure of uncertainty

- *Probability is a measure of certainty of an event taking place.*
- *i.e. in the example, we were measuring the chances of hitting the shaded area.*



Its area is 1

$$prob = \frac{\#\,\mathrm{Re}\,dBoxes}{\#\,Boxes}$$

Adapt from Prof. Nando de Freitas's review slides

Today : Probability Review

- The big picture
- Sample space, Event and Event spaces
- Random variables
- Joint probability, Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
- Structural properties
  - Independence, conditional independence

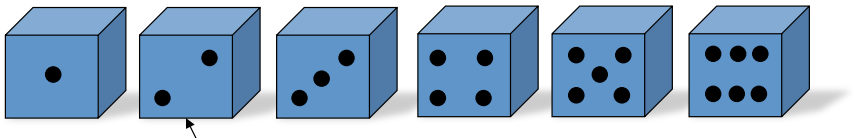
Probability

Probability *is the formal study of the laws of chance. Probability allows us to* ***manage uncertainty****.*

The ***sample space*** *is the set of all* ***outcomes****. For example, for a die we have 6 outcomes:*

$$O_{die} = \{1,2,3,4,5,6\}$$

O:



Elementary Event "Throw a 2"

The elements of O are called *elementary events.*

Probability

- *Probability allows us to measure many **events**.*
- ***The events are subsets of the sample space O**. For example, for a die we may consider the following events: e.g.,*

GREATER = {5, 6}

EVEN = {2, 4, 6}

- ***Assign probabilities to these events: e.g.,***

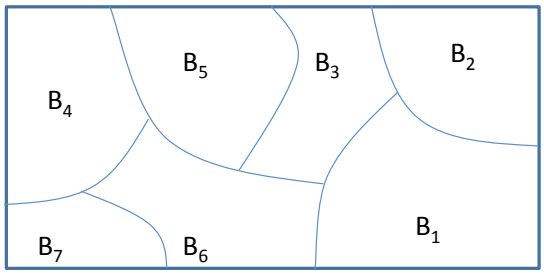
P(EVEN) = 1/2

Sample space and Events

- W : Sample Space, result of an experiment
  - If you toss a coin twice W= {HH,HT,TH,TT}
- Event: a subset of W
  - First toss is head = {HH,HT}
- S: event space, a set of events:
  - Contains the empty event and W

Axioms for Probability

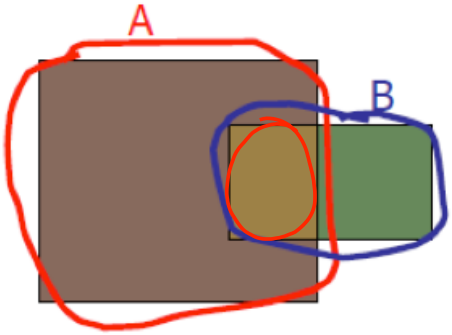
- Defined over (W,S) s.t.
  - 1 >= P(a) >= 0 for all a in S
  - P(W) = 1
  - If A, B are disjoint, then
    - P(A U B) = p(A) + p(B)
- P(W) = $\sum P(B_i)$



OR operation for Probability

- We can deduce other axioms from the above ones
  - Ex: P(A U B) for non-disjoint events

P(A or B) = P(A) + P(B) - P(A and B)



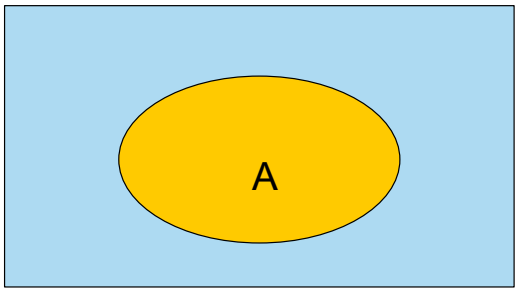
Theorems from the Axioms

- 0 <= P(A) <= 1,
- P(A or B) = P(A) + P(B) - P(A and B)

From these we can prove:

P(not A) = P(~A) = 1-P(A)

Complement



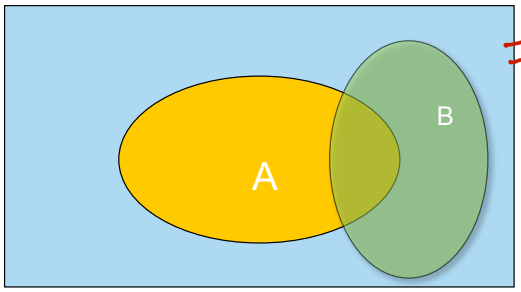
Another important theorem

- 0 <= P(A) <= 1,
- P(A or B) = P(A) + P(B) - P(A and B)

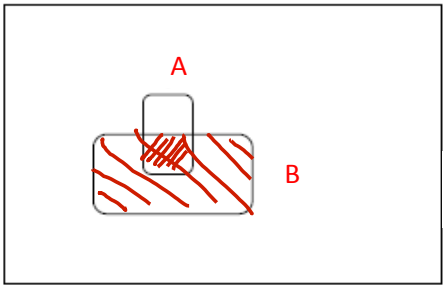
From these we can prove:

P(A) = P(A ^ B) + P(A ^ ~B)

$P(A)$
$= P(A \wedge \Omega)$
$= P(A \wedge (B \vee \neg B))$
$= P((A \wedge B) \vee (A \wedge \neg B))$
$= P(A \wedge B) + P(A \wedge \neg B)$



Conditional Probability



$$P(A \text{ given } B) = P(A \text{ and } B) \,/\, P(B)$$

That is, in the frequentist interpretation, we calculate the ratio of the number of times both A and B occurred and divide it by the number of times B occurred.

For short we write: $P(A|B) = P(AB)/P(B)$; *or* $P(AB)=P(A|B)P(B)$, *where* $P(A|B)$ *is the* ***conditional*** *probability,* $P(AB)$ *is the* ***joint****, and* $P(B)$ *is the* ***marginal****.*

If we have more events, we use the chain rule:

$$P(ABC) = P(A|BC)\,P(B|C)\,P(C)$$

$P(ABC) = P(A|BC)P(BC) = P(A|BC)P(B|C)P(C)$

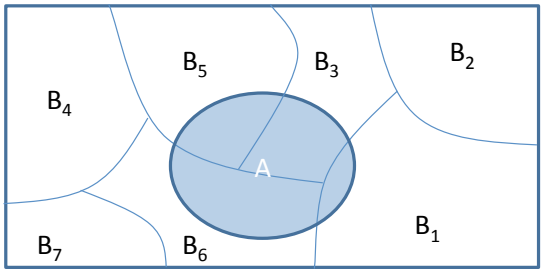
from Prof. Nando de Freitas's review

Conditional Probability / Chain Rule

- More ways to write out chain rule …

$$P(A,B) = p(B|A)p(A)$$
$$P(A,B) = p(A|B)p(B)$$

Rule of total probability => Marginalization



$$\Rightarrow p(A) = \sum P(B_i) P(A \mid B_i)$$

WHY ???

$P(A) = P(A \cap \Omega)$

$= P(A \cap (B_1 \cup B_2 \cdots \cup B_k))$

$= \sum P(A \cap B_i)$

Today : Probability Review

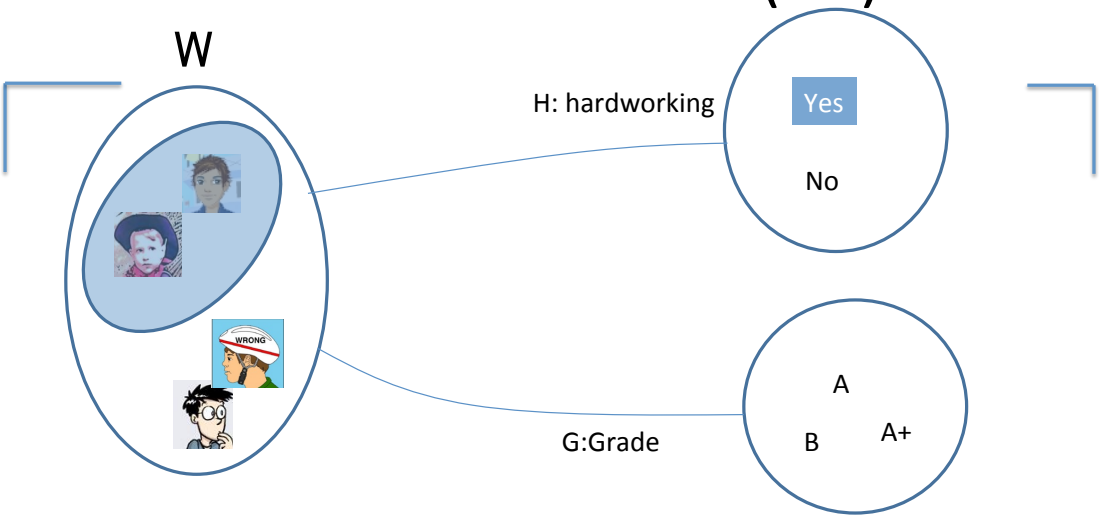
- The big picture
- Events and Event spaces
- Random variables
- Joint probability, Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
- Structural properties
  - Independence, conditional independence

From Events to Random Variable

- Concise way of specifying attributes of outcomes
- Modeling students (Grade and Intelligence):
  - W= all possible students (sample space)
  - What are events (subset of sample space)
    - Grade_A = all students with grade A
    - Grade_B = all students with grade B
    - HardWorking_Yes = … who works hard
  - Very cumbersome

  - Need "functions" that maps from W to an attribute space T.
  - P(H = YES) = P({student ϵ W : H(student) = YES})

Random Variables (RV)



P(H = Yes) = P({all students who is working hard on the course})

Notation Digression

- P(A) is shorthand for P(A=true)
- P(~A) is shorthand for P(A=false)
- Same notation applies to other binary RVs: P(Gender=M), P(Gender=F)
- Same notation applies to *multivalued* RVs: P(Major=history), P(Age=19), P(Q=c)
- Note: upper case letters/names for *variables*, lower case letters/names for *values*

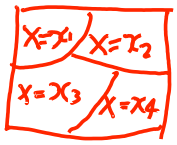
Discrete Random Variables

- Random variables (RVs) which may take on only a **countable** number of **distinct** values

- X is a RV with arity *k* if it can take on exactly one value out of $\{x_1, \ldots, x_k\}$

Probability of Discrete RV

- Probability mass function (pmf): $P(X = x_i)$
- Easy facts about pmf
  - $\Sigma_i P(X = x_i) = 1$
  - $P(X = x_i \cap X = x_j) = 0$ if $i \neq j$
  - $P(X = x_i \cup X = x_j) = P(X = x_i) + P(X = x_j)$ if $i \neq j$
  - $P(X = x_1 \cup X = x_2 \cup \ldots \cup X = x_k) = 1$



e.g. Coin Flips

- You flip a coin
  - Head with probability 0.5
- You flip 100 coins
  - How many heads would you expect

e.g. Coin Flips cont.

- You flip a coin
  - Head with probability *p*
  - Binary random variable
  - Bernoulli trial with success probability *p*
- You flip *k* coins
  - How many heads would you expect
  - Number of heads X: discrete random variable
  - Binomial distribution with parameters *k* and *p*

Discrete Random Variables

- Random variables (RVs) which may take on only a **countable** number of **distinct** values
  - E.g. the total number of heads X you get if you flip 100 coins
- X is a RV with arity *k* if it can take on exactly one value out of $\{x_1, \ldots, x_k\}$
  - E.g. the possible values that X can take on are 0, 1, 2,…, 100

e.g., two Common Distributions

- Uniform $X \sim U[1,...,N]$
  - X takes values 1, 2, …, *N*
  - $P(X=i)=1/N$
  - E.g. picking balls of different colors from a box

- Binomial $X \sim Bin(k,p)$
  - X takes values 0, 1, …, *k*
  - $P(X=i)=\binom{k}{i} p^i (1-p)^{k-i}$
  - E.g. coin flips k times

Today : Probability Review

- The big picture
- Events and Event spaces
- Random variables
- Joint probability, Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
- Structural properties
  - Independence, conditional independence

e.g., Coin Flips by Two Persons

- Your friend and you both flip coins
  - Head with probability 0.5
  - You flip 50 times; your friend flip 100 times
  - How many heads will both of you get

Joint Distribution

- Given two discrete RVs X and Y, their **joint distribution** is the distribution of X and Y together $P((X=21) \cap (Y=70))$
  - E.g. P(You get 21 heads AND you friend get 70 heads)
-
  - E.g. sum $\sum_x \sum_y P(X=x \cap Y=y) = 1$

$$\sum_{i=0}^{50} \sum_{j=0}^{100} P(\text{You get } i \text{ heads AND your friend get } j \text{ heads}) = 1$$

Conditional Probability

- $P(X = x | Y = y)$ is the probability of $X = x$, given the occurrence of $Y = y$
  - E.g. you get 0 heads, given that your friend gets 61 heads
- $$P(X = x | Y = y) = \frac{P(X = x \cap Y = y)}{P(Y = y)}$$

Law of Total Probability

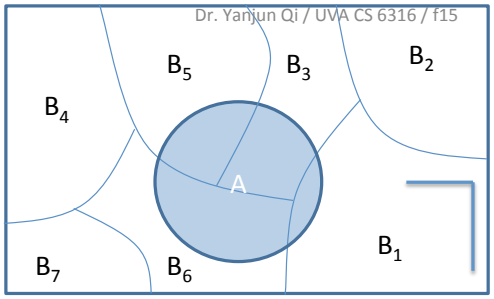
- Given two discrete RVs X and Y, which take values in $\{x_1, \ldots, x_m\}$ and $\{y_1, \ldots, y_n\}$, We have

$$= P(X = x_i \cap \Omega)$$

$$P(X = x_i) = \sum_j P(X = x_i \cap Y = y_j)$$

$$= \sum_j P(X = x_i | Y = y_j) P(Y = y_j)$$

Marginalization



Marginal Probability

Joint Probability

$$P(X = x_i) = \sum_j P(X = x_i \cap Y = y_j)$$

$$= \sum_j P(X = x_i | Y = y_j) P(Y = y_j)$$

chain rule

Conditional Probability

Marginal Probability

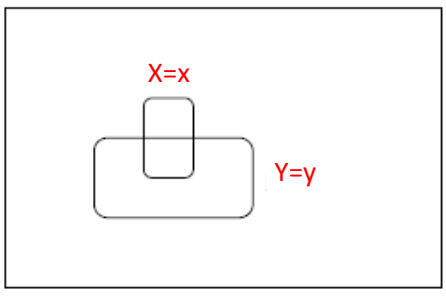
Simplify Notation: Conditional Probability

events

$$P(X = x | Y = y) = \frac{P(X = x \cap Y = y)}{P(Y = y)}$$

But we will always write it this way:

$$P(x | y) = \frac{p(x, y)}{p(y)}$$

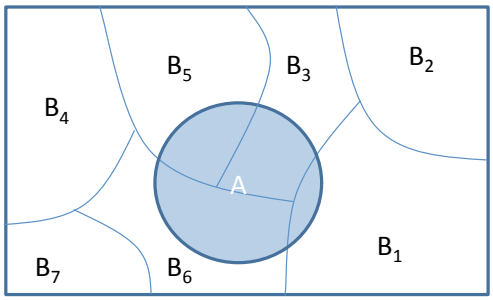


Simplify Notation: Marginalization

- We know p(X, Y), what is P(X=x)?
- We can use the law of total probability, why?

$$p(x) = \sum_y P(x, y)$$

$$= \sum_y P(y)P(x \mid y)$$



Marginalization Cont.

- Another example

$$p(x) = \sum_{y,z} P(x, y, z)$$

$$= \sum_{z,y} P(y, z)P(x \mid y, z)$$

Bayes Rule

- We know that P(rain) = 0.5
  - If we also know that the grass is wet, then how this affects our belief about whether it rains or not?

$$P(rain \mid wet) = \frac{P(rain)P(wet \mid rain)}{P(wet)}$$

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

Bayes Rule

- X and Y are discrete RVs…

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

$$P(X = x_i \mid Y = y_j) = \frac{P(Y = y_j \mid X = x_i)P(X = x_i)}{\sum_k P(Y = y_j \mid X = x_k)P(X = x_k)}$$

What we just did…

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

This is Bayes Rule

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

More General Forms of Bayes Rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\sim A)P(\sim A)}$$

$$P(A|B \wedge X) = \frac{P(B|A \wedge X)P(A \wedge X)}{P(B \wedge X)}$$

$$P(A = a_1 \mid B) = \frac{P(B \mid A = a_1)P(A = a_1)}{\sum_i P(B \mid A = a_i)P(A = a_i)}$$

Bayes Rule cont.

- You can condition on more variables

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

Conditional Probability Example

Assume we have a dark box with 3 red balls and 1 blue ball. That is, we have the set {r,r,r,b}. What is the probability of drawing 2 red balls in the first 2 tries?

$$P(B_1 = r, B_2 = r) = P(B_1 = r)\, P(B_2 = r \mid B_1 = r)$$

$$= \frac{3}{4} \times \frac{2}{3} = \frac{1}{2}$$

Adapt from Prof. Nando de Freitas's review slides

Conditional Probability Example

What is the probability that the 2nd ball drawn from the ***set*** *{r,r,r,b} will be red?*

Using marginalization, $P(B_2 = r) = P(B_2 = r \wedge B_1 = r) + P(B_2 = r \wedge B_1 = b)$

$= P(B_1 = r) P(B_2 = r \mid B_1 = r) + P(B_1 = b) P(B_2 = r \mid B_1 = b)$

$$\begin{bmatrix} P(B_2 = r) \\ P(B_2 = b) \end{bmatrix} = \begin{bmatrix} P(B_2 = r \mid B_1 = r) P(B_1 = r) + P(B_2 = r \mid B_1 = b) P(B_1 = b) \\ P(B_2 = b \mid B_1 = r) P(B_1 = r) + P(B_2 = b \mid B_1 = b) P(B_1 = b) \end{bmatrix}$$

$$= \begin{bmatrix} P(B_2 = r \mid B_1 = r), & P(B_2 = r \mid B_1 = b) \\ P(B_2 = b \mid B_1 = r), & P(B_2 = b \mid B_1 = b) \end{bmatrix} \begin{bmatrix} P(B_1 = r) \\ P(B_1 = b) \end{bmatrix}$$

$P(B_2)$ $\quad P(B_2 \mid B_1)$ $\quad P(B_1)$

➔ Matrix Notation

$\Rightarrow$ matrix notation form $\Pi_2 = G^T \Pi_1$

For short, we write this using vectors and a ***stochastic matrix****:*

Today : Probability Review

- The big picture
- Events and Event spaces
- Random variables
- Joint probability, Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
- Structural properties
  - Independence, conditional independence

Next

Today Recap : Probability Review

- The big picture
- Events and Event spaces
- Random variables
- Joint probability, Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.

References

- Prof. Andrew Moore's review tutorial
- Prof. Nando de Freitas's review slides
- Prof. Carlos Guestrin recitation slides