Alex Smola

Barnabas Poczos

TA: Ina Fiterau

4th year PhD student MLD

Review of Probabilities and Basic Statistics

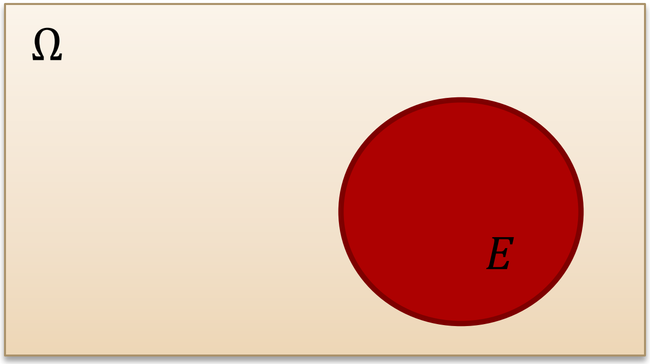
10-701 Recitations

Overview

- Introduction to Probability Theory
- Random Variables. Independent RVs
- Properties of Common Distributions
- Estimators. Unbiased estimators. Risk
- Conditional Probabilities/Independence
- Bayes Rule and Probabilistic Inference

Review: the concept of probability

- Sample space Ω – set of all possible outcomes
- Event $\mathrm{E} \in \Omega$ – a subset of the sample space
- Probability measure – maps Ω to unit interval
  - "How likely is that event E will occur?"
- Kolmogorov axioms
  - $\mathrm{P}(\mathrm{E}) \geq 0$
  - $\mathrm{P}(\Omega) = 1$
  - $\mathrm{P}(E_1 \cup E_2 \cup \cdots) = \sum_{i=1}^{\infty} P(E_i)$



Reasoning with events

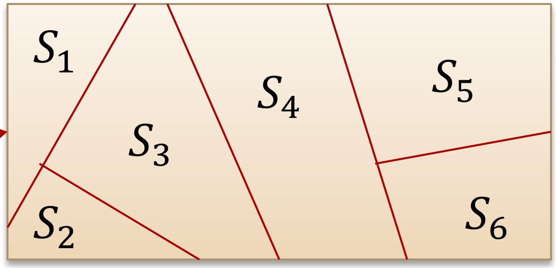
- Venn Diagrams
  - $P(A) = Vol(A)/Vol\,(\Omega)$
- Event union and intersection
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Properties of event union/intersection
  - Commutativity: $A \cup B = B \cup A$; $A \cap B = B \cap A$
  - Associativity: $A \cup (B \cup \mathrm{C}) = (A \cup B) \cup \mathrm{C}$
  - Distributivity: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

Reasoning with events

- DeMorgan's Laws
  - $(A \cup B)^C = A^C \cap B^C$
  - $(A \cap B)^C = A^C \cup B^C$
- Proof for law #1 - by double containment
  - $(A \cup B)^C \subseteq A^C \cap B^C$
    - …
  - $A^C \cap B^C \subseteq (A \cup B)^C$
    - …

Reasoning with events

- Disjoint (mutually exclusive) events
  - $P(A \cap B) = 0$
  - $P(A \cup B) = P(A) + P(B)$
  - examples:
    - A and A^C
    - partitions



- NOT the same as independent events
  - For instance, successive coin flips

Partitions

- Partition $S_1 \ldots S_n$
  - Events cover sample space $S_1 \cup \cdots \cup S_n = \Omega$
  - Events are pairwise disjoint $S_i \cap S_j = \emptyset$
- Event reconstruction
  - $P(A) = \sum_{i=1}^{n} P(A \cap S_i)$
- Boole's inequality
  - $P(\cup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{n} P(A_i)$
- Bayes' Rule
  - $P(S_i|A) = \frac{P(A|S_i)P(S_i)}{\sum_{j=1}^{n} P(A|S_j)P(S_j)}$

Overview

- Introduction to Probability Theory
- Random Variables. Independent RVs
- Properties of Common Distributions
- Estimators. Unbiased estimators. Risk
- Conditional Probabilities/Independence
- Bayes Rule and Probabilistic Inference

Random Variables

- Random variable – associates a value to the outcome of a randomized event
- Sample space $\mathcal{X}$: possible values of rv X
- Example: event to random variable

Draw 2 numbers between 1 and 4. Let r.v. X be their sum.																
E	11	12	13	14	21	22	23	24	31	32	33	34	41	42	43	44
X(E)	2	3	4	5	3	4	5	6	4	5	6	7	5	6	7	8

Induced probability function on $\mathcal{X}$.							
x	2	3	4	5	6	7	8
P(X=x)	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

Cumulative Distribution Functions

- $F_X(x) = P(X \leq x) \; \forall x \in \mathcal{X}$
- The CDF completely determines the probability distribution of an RV
- The function $F(x)$ is a CDF i.i.f
  - $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$
  - $F(x)$ is a non-decreasing function of x
  - $F(x)$ is right continuous: $\forall x_0 \quad \lim_{\substack{x \to x_0 \\ x > x_0}} F(x) = F(x_0)$

Identically distributed RVs

- Two random variables X_1 and X_2 are identically distributed iif for all sets of values A

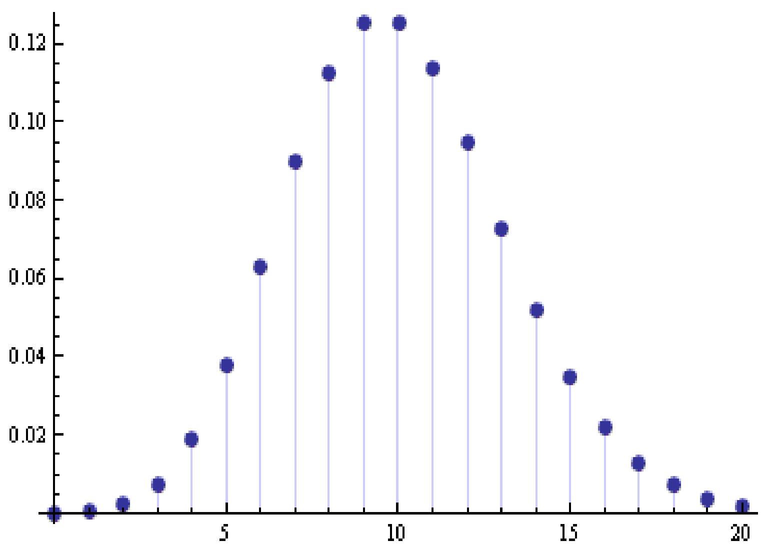
$$P(X_1 \in A) = P(X_2 \in A)$$

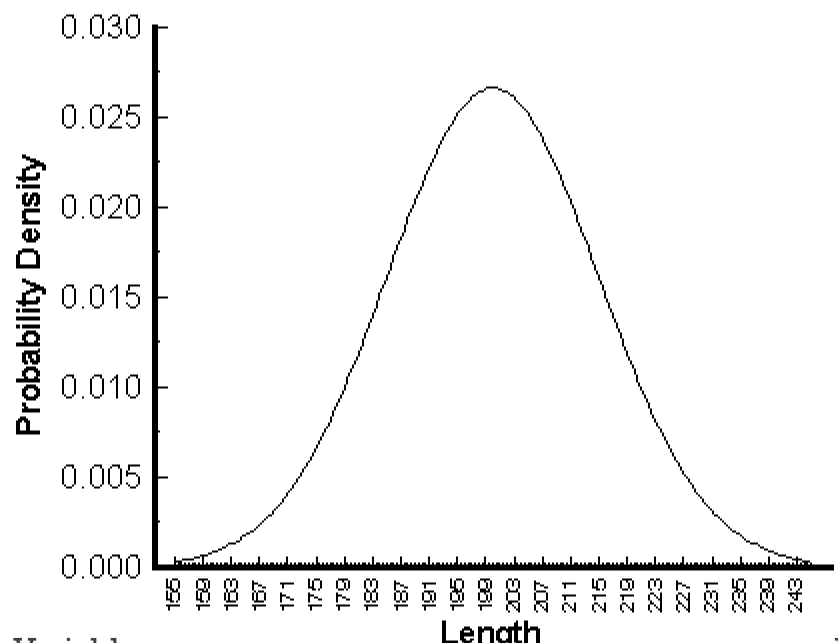
- So that means the variables are equal?
  - NO.
  - Example: Let's toss a coin 3 times and let X_H and X_F represent the number of heads/tails respectively
  - They have the same distribution but $X_H = 1 - X_F$

Discrete vs. Continuous RVs

- Step CDF
- $\mathcal{X}$ is discrete
- Probability mass
  - $f_X(x) = P(X = x) \ \ \forall x$

- Continuous CDF
- $\mathcal{X}$ is continuous
- Probability density
  - $F_X(x) = \int_{-\infty}^{x} f_X(t)dt \ \ \forall x$

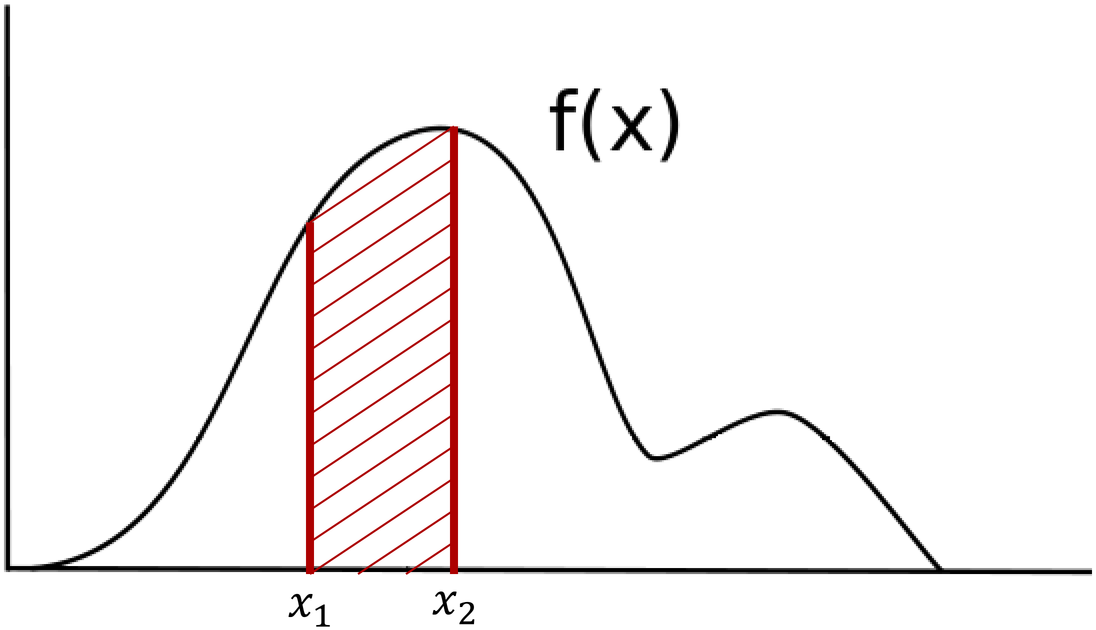




Interval Probabilities

- Obtained by integrating the area under the curve

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f_x(x)dx$$



- This explains why P(X=x) = 0 for continuous distributions!

$$P(X = x) \leq \lim_{\substack{\epsilon \to 0 \\ \epsilon > 0}} [F_x(x) - F_x(x - \epsilon)] = 0$$

Moments

- Expectations
  - The expected value of a function g depending on a r.v. $X \sim P$ is defined as $Eg(X) = \int g(x)P(x)dx$
- n$^{\text{th}}$ moment of a probability distribution

$$\mu_n = \int x^n P(x)dx$$

- mean $\mu = \mu_1$
- n$^{\text{th}}$ central moment

$$\mu_n{}' = \int (x-\mu)^n P(x)dx$$

- Variance $\sigma^2 = \mu_2{}'$

Multivariate Distributions

- Example
  - Uniformly draw X and Y from the set $\{1,2,3\}^2$
  - $W = X + Y; V = |X - Y|$

W \ V	0	1	2	P_W
2	1/9	0	0	1/9
3	0	2/9	0	2/9
4	1/9	0	2/9	3/9
5	0	2/9	0	2/9
6	1/9	0	0	1/9
P_V	3/9	4/9	2/9	1

- Joint
  - $P\big((X,Y) \in A\big) = \sum_{(x,y)\epsilon A} f(x,y)$
- Marginal
  - $f_Y(y) = \sum_x f(x,y)$
- For independent RVs:
  - $f(x_1, \ldots, x_n) = f_{X_1}(x_1) \ldots f_{X_n}(x_n)$

Overview

- Introduction to Probability Theory
- Random Variables. Independent RVs
- Properties of Common Distributions
- Estimators. Unbiased estimators. Risk
- Conditional Probabilities/Independence
- Bayes Rule and Probabilistic Inference

Bernoulli

- $X = \begin{cases} 1 & with\ probability\ p \\ 0 & with\ probability\ 1-p \end{cases} \quad 0 \le p \le 1$
- Mean and Variance
  - $EX = 1p + 0(1-p) = p$
  - $VarX = (1-p^2)p + (0-p^2)(1-p) = p(1-p)$
- MLE: sample mean
- Connections to other distributions:
  - If $X_1 \ldots X_n \sim Bern(p)$ then $\mathrm{Y} = \sum_{i=1}^{n} X_i$ is $\mathrm{Binomial(n,p)}$
  - Geometric distribution – the number of Bernoulli trials needed to get one success

Binomial

- $P(X = x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}$
- Mean and Variance
  - $EX = \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x} = \ldots = np$
  - $VarX = np(1-p)$
  - NOTE: $\boldsymbol{VarX = EX^2 - (EX)^2}$
- Sum of Bin is Bin
- Conditionals on Bin are Bin

Properties of the Normal Distribution

- Operations on normally-distributed variables
  - $X_1, X_2 \sim Norm(0,1)$, then $X_1 \pm X_2 \sim N(0,2)$
  - $X_1/X_2 \sim Cauchy(0,1)$
  - $X_1 \sim Norm(\mu_1, {\sigma_1}^2), X_2 \sim Norm(\mu_2, {\sigma_2}^2)$ and $X_1 \perp X_2$

  then $Z = X_1 + X_2 \sim Norm(\mu_1 + \mu_2, {\sigma_1}^2 + {\sigma_2}^2)$

- If $X , Y \sim N\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} {\sigma_X}^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & {\sigma_Y}^2 \end{pmatrix}\right)$, then

$X + Y$ is still normally distributed, the mean is the sum of the means and the variance is

${\sigma_{X+Y}}^2 = {\sigma_X}^2 + {\sigma_Y}^2 + 2\rho\sigma_X\sigma_Y$, where ρ is the correlation

Overview

- Introduction to Probability Theory
- Random Variables. Independent RVs
- Properties of Common Distributions
- Estimators. Unbiased estimators. Risk
- Conditional Probabilities/Independence
- Bayes Rule and Probabilistic Inference

Estimating Distribution Parameters

- Let $X_1 \ldots X_n$ be a sample from a distribution parameterized by θ
- How can we estimate
  - The mean of the distribution?
  - Possible estimator: $\frac{1}{n}\sum_{i=1}^{n} X_i$
  - The median of the distribution?
  - Possible estimator: $median(X_1 \ldots X_n)$
  - The variance of the distribution?
  - Possible estimator: $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$

Bias-Variance Tradeoff

- When estimating a quantity θ, we evaluate the performance of an estimator by computing its risk – expected value of a loss function
  - $\mathrm{R}\left(\theta, \hat{\theta}\right) = E\ L(\theta, \hat{\theta})$, where L could be
    - Mean Squared Error Loss
    - 0/1 Loss
    - Hinge Loss (used for SVMs)
- Bias-Variance Decomposition: $Y = f(x) + \varepsilon$

$$Err(x) = E\left[f(x) - \hat{f}(x)^2\right]$$

$$= \underbrace{(E\left[\hat{f}(x)\right] - f(x))^2}_{\text{Bias}} + \underbrace{E\left[\hat{f}(x) - E\left[\hat{f}(x)\right]\right]^2}_{\text{Variance}} + {\sigma_\varepsilon}^2$$

Overview

- Introduction to Probability Theory
- Random Variables. Independent RVs
- Properties of Common Distributions
- Estimators. Unbiased estimators. Risk
- Conditional Probabilities/Independence
- Bayes Rule and Probabilistic Inference

Review: Conditionals

- Conditional Variables
  - $P(X|Y) = \frac{P(X,Y)}{P(Y)}$ note X;Y is a different r.v.
- Conditional Independence $X \perp Y \,|Z$
  - X and Y are cond. independent given Z iif
    $$P((X,Y)|Z) = P(X|Z)P(Y|Z)$$
- Properties of Conditional Independence
  - Symmetry $X \perp Y \,|Z \;\Leftrightarrow Y \perp X \,|Z$
  - Decomposition $X \perp (Y,W)|Z \;\Rightarrow X \perp Y \,|Z$
  - Weak Union $X \perp (Y,W)|Z \;\Rightarrow X \perp Y \,|Z,W$
  - Contraction $(X \perp W|Z,Y)\,,(X \perp Y \,|Z) \Rightarrow X \perp Y,W \,|Z$

Can you prove these?

Overview

- Introduction to Probability Theory
- Random Variables. Independent RVs
- Properties of Common Distributions
- Estimators. Unbiased estimators. Risk
- Conditional Probabilities/Independence
- Bayes Rule and Probabilistic Inference

Priors and Posteriors

- We've so far introduced the likelihood function
  - $P(Data|\theta)$ - the likelihood of the data given the parameter of the distribution
  - $\theta_{MLE} = argmax_\theta P(Data|\theta)$
- What if not all values of θ are equally likely?
  - θ itself is distributed according to the prior P_θ
  - Apply Bayes rule
    - $P(\theta|Data) = \frac{P(Data|\theta)P(\theta)}{P(Data)}$
    - $\theta_{MAP} = argmax_\theta P(\theta|Data) = argmax_\theta P(Data|\theta)P(\theta)$

Conjugate Priors

- If the posterior distributions $P(\theta|Data)$ are in the **same family** as the prior prob. distribution P_θ, then the prior and the posterior are called **conjugate distributions** and P_θ is called **conjugate prior**
- Some examples

Likelihood	Conjugate Prior
Bernoulli/Binomial	Beta
Poisson	Gamma
(MV) Normal with known (co)variance	Normal
Exponential	Gamma
Multinomial	Dirichlet

How to compute the parameters of the Posterior?

I'll send a derivation

Probabilistic Inference

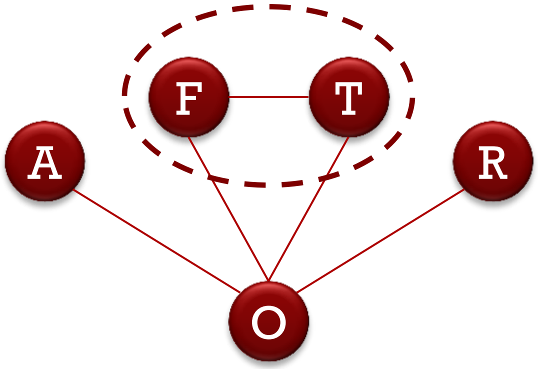
- Problem: You're planning a weekend biking trip with your best friend, Min. Alas, your path to outdoor leisure is strewn with many hurdles. If it happens to rain, your chances of biking reduce to half not counting other factors. Independent of this, Min might be able to bring a tent, the lack of which will only matter if you notice the symptoms of a flu before the trip. Finally, the trip won't happen if your advisor is unhappy with your weekly progress report.

Probabilistic Inference

- Problem: You're planning a weekend biking trip with your best friend, Min. Your path to outdoor leisure is strewn with many hurdles. If it happens to rain, your chances of biking reduce to half not counting other factors. Independent of this, Min might be able to bring a tent, the lack of which will only matter if you notice the symptoms of a flu before the trip. Finally, the trip won't happen if your advisor is unhappy with your weekly progress report.
- Variables:
  - O – the outdoor trip happens
  - A – advisor is happy
  - R – it rains that day
  - T – you have a tent
  - F – you show flu symptoms

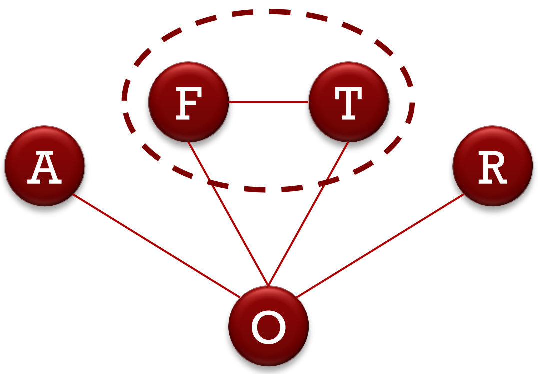
Probabilistic Inference

- Problem: You're planning a weekend biking trip with your best friend, Min. Alas, your path to outdoor leisure is strewn with many hurdles. If it happens to rain, your chances of biking reduce to half not counting other factors. Independent of this, Min might be able to bring a tent, the lack of which will only matter if you notice the symptoms of a flu before the trip. Finally, the trip won't happen if your advisor is unhappy with your weekly progress report.
- Variables:
  - O – the outdoor trip happens
  - A – advisor is happy
  - R – it rains that day
  - T – you have a tent
  - F – you show flu symptoms



Probabilistic Inference

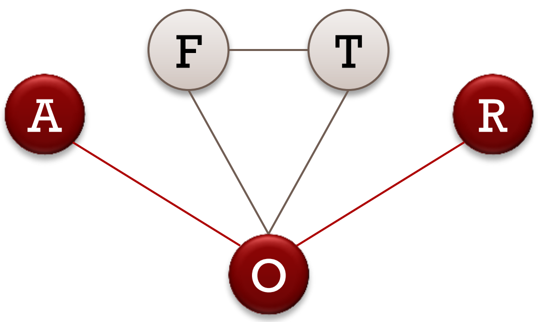
- How many parameters determine this model?
- P(A|O) => 1 parameter
- P(R|O) => 1 parameter
- P(F, T|O) => 3 parameters
- In this problem, the values are given;
- Otherwise, we would have had to estimate them
- Variables:
  - O – the outdoor trip happens
  - A – advisor is happy
  - R – it rains that day
  - T – you have a tent
  - F – you show flu symptoms



Probabilistic Inference

- The weather forecast is optimistic, the chances of rain are 20%. You've barely slacked off this week so your advisor is probably happy, let's give it an 80%. Luckily, you don't seem to have the flu.
- What are the chances that the trip will happen?

Think of how you would do this.
Hint #1: do the variables F and T influence the result in this case?
Hint #2: use the fact that the combinations of values for A and R represent a partition and use one of the partition formulas we learned



Overview

- Introduction to Probability Theory
- Random Variables. Independent RVs
- Properties of Common Distributions
- Estimators. Unbiased estimators. Risk
- Conditional Probabilities/Independence
- Bayes Rule and Probabilistic Inference

Overview

- Introduction to Probability Theory
- Random Variables. Independent RVs
- Properties of Common Distributions
- Estimators. Unbiased estimators. Risk
- Conditional Probabilities/Independence
- Bayes Rule and Probabilistic Inference

Questions?