B555 2018

Probability Read Wasserman Ch 1

Probability deals with exps. having random outcomes

Ex

① Flip a coin

② Roll a die

③ Sample a voter

Defn

The sample space, Ω, is set of all possible outcomes

Ex

Flip 2 coins: $\Omega = \{HH, HT, TH, TT\}$

Choose card from deck $\Omega = \{A, 2, \dots 10, J, Q, K\} \times \{\clubsuit, \diamondsuit, \heartsuit, \spadesuit\}$

Defn

An event is a subset of ~~pass~~ Ω.

Ex

A = get a head in 2 coin flips

$A = \{HT, TH, HH\} \subseteq \Omega$

Since events are subsets usual set operations are possible

A^c = A did not occur $A \cup B$: Either A, B, or both occured.

$A \cap B$: Both A and B occurred [$\emptyset$ and Ω also events]

Defn

A probability is a set function P, $P(A)$, for all possible subsets satisfying

① $P(A) \geq 0$

② $P(\Omega) = 1$

③ For disjoint $A_1 \cdots A_n$ $(A_i \cap A_j = \emptyset \text{ for } i \neq j)$ $P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$

and

For disjoint $A_1, A_2, \cdots$ $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

Simple relations follow

Ex

$P(A^c) = 1 - P(A)$

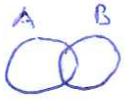
$[\ 1 = P(\Omega) = P(A \cup A^c) = P(A) + P(A^c)\ ]$

Ex

$P(A \cup B) = P(A) + P(B) - P(A, B)$

$[\ P(A \cap B)\ ]$

PF



$P(A \cup B) = \big(P(A \cap B^c) + P(A \cap B)\big) + \big(P(A^c \cap B) + P(A \cap B)\big) - P(A \cap B)$

$= P(A) + P(B) - P(A \cap B)$

Equally Likely Outcomes

Often reasonable to assume all outcomes = 'Equally likely'

$\Omega = \{HH, HT, TH, TT\}$. Then $P(A) = \frac{|A|}{|\Omega|}$

Ex Roll 2 dice

$\Omega = \{1, 2, \cdots 6\} \times \{1, 2, \cdots 6\}$

36 "Equally" outcomes each where has prob. $1/36$.

$P(\text{Sum} = 7) = \frac{|\text{Sum} = 7|}{|\Omega|} = \frac{6}{36}$

Independence

Defn

A, B independent if $P(A,B) = P(A)P(B)$

$A_1 \dots A_n$ mutually independent if for every subset $J \subseteq \{1, \dots n\}$

$$P\left(\bigcap_{j \in J} A_j\right) = \prod_{j \in J} P(A_j)$$

Intuitively if A has no effect on B and vice versa

A occurs $P(A)$ of time

Of times when A occurs, B occurs $P(B)$ of time

So A and B occur $P(A)P(B)$ of time.

That is $P(A,B) = P(A)P(B)$.

Ex

Toss coin 10 times. What is $P(\text{at least } 1 \text{ H})$?

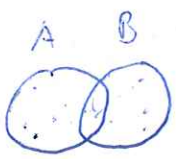
$$P(\text{At least } 1 \text{ H}) = 1 - P(\text{No heads})$$

$$= 1 - P(1^{st} \text{ flip T}, \dots 10^{th} \text{ flip T})$$

$$= 1 = 1 - \left(\frac{1}{2}\right)^{10}$$

Conditional Probability

$P(A|B)$ is prob of A occurring when we know B occurs.



$P(A|B)$ = "prob of A given B"

Defn If $P(B) > 0$ $P(A|B) = \frac{P(A,B)}{P(B)}$

Note from Defn $P(A,B) = P(A|B)P(B)$

Ex Have 2 white + 2 black balls in urn. Draw 2 balls w/o replacement. What is prob both black?

$$P(\text{Both black}) = P(B_1 \text{ black}, B_2 \text{ black}) = P(B_2 \text{ black} \mid B_1 \text{ black})\, P(B_1 \text{ black})$$

$$= \left(\frac{1}{3}\right)\left(\frac{1}{2}\right) = \frac{1}{6}.$$

Ex Law of Rare disease

Have rare disease existing in $\frac{1}{1000}$ of population

Have test for disease s.t.

Test gives + result when person has disease w/ prob. .99

" " − " " " doesn't have " " " "

Randomly chosen person tests positively for disease. What is prob. person has disease?

Let D = person has disease $\quad P(D) = .001$

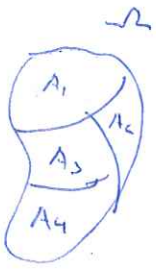
$P(D^c) = .999$

Have $P(+ \mid D) = P(- \mid D^c) = .99$

$$P(D \mid +) = \frac{P(D, +)}{P(+)} = \frac{P(D)\,P(+ \mid D)}{P(D)\,P(+ \mid D) + P(D^c)\,P(+ \mid D^c)}$$

$$P(+) = P((+, D) \cup (+, D^c)) = \frac{(.001)(.99)}{(.001)(.99) + (.999)(.01)} \approx \frac{1}{11}$$

Law of Total Prob



Let $A_1 \cdots A_n$ be partition of Ω

$\left(A_i \cap A_j = \emptyset \quad i \neq j \; ; \quad \dot{\bigcup}_{i=1}^{n} A_i = \Omega \right)$

Then $P(B) = \sum_{i=1}^{n} P(A_i) P(B|A_i)$

Pf

$B = \dot{\bigcup}_{i=1}^{n} A_i \cap B$ and $A_1 \cap B$, $A_2 \cap B$ ... $A_n \cap B$ are disjoint so

$$P(B) = \sum_{i=1}^{n} P(A_i \cap B) = \sum_{i=1}^{n} P(A_i) P(B|A_i)$$

Ex

Have 3 black, 4 white, 2 red socks in drawer.

What is prob of choosing a matching pair?

A_1 = 1st sock black [note A_1, A_2, A_3 partition Ω]

A_2 = " " white B = matching pair

A_3 = " " red

$$P(B) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3)$$

$$= \frac{3}{9}\frac{2}{8} + \frac{4}{9}\frac{3}{8} + \frac{2}{9}\frac{1}{8} = \frac{15}{72}$$

Bayes' Rule

Let $A_1 \cdots A_n$ partition Ω. ~~Then~~

Suppose we know $P(B|A_i)$ $i = 1 \cdots n$.

Then $P(A_i | B) = \frac{P(A_i, B)}{P(B)} = \frac{P(A_i) P(B|A_i)}{\sum_{j=1}^{n} P(A_j) P(B|A_j)}$

Random Variables Rest Wiseman Ch2

Defn

A Random variable (rv) is function from Ω to $\mathbb{R}$ (Real #'s)

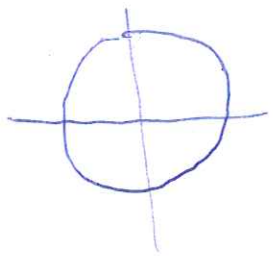
Ex

Flip coin 3 times sc $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$

X = # H's $\quad$ 3 $\quad$ 2 $\quad$ 2 $\quad$ ...

Ex

Choose a point randomly from unit disc $= \{(x,y) \in \mathbb{R}^2 : x^2+y^2 \leq 1\}$



Let W = dist. from origin

$W = (x^2+y^2)^{1/2}$

Defns

A discrete r.v. has enumerable outcomes $\Omega = \{\omega_1, \ldots \omega_n\}$

or $\Omega = \{\omega_1, \omega_2, \ldots\}$

A continuous r.v. has one or several intervals of $\mathbb{R}$ as possible outcomes

$[0,1]$, $\mathbb{R}^{\#}$, $[0,\infty)$, $[0, \frac{1}{2}) \cup [1, \frac{3}{2})$

Important Discrete Distributions

Bernoulli

If $X \in \{0, 1\}$ with $P(X=1) = p$
$P(X=0) = q = 1-p$

X is said to be Bernoulli. Often written

$$f(x) = p^x (1-p)^{1-x} \quad x = 0, 1$$

Bernoulli is # of successes from single S-F expt. $P(S) = p$.

Binomial

Have c indep. S-F trials with $P(S) = p$ $P(F) = 1-p$
Let X = # of S's
$f(x) = \binom{c}{x} p^x (1-p)^{c-x}$ $x = 0, \ldots, c$

Geometric

Have S-F expt. $P(S) = p$. Perform repeated S-F expts until 1st S.
Let X = # trials until S.

$P(X=1) = P(S) = p$
$P(X=2) = P(FS) = (1-p)p$
$P(X=3) = P(FFS) = (1-p)^2 p$
$\vdots$
$P(X=x) = P(\underbrace{FF\cdots F}_{x-1}S) = (1-p)^{x-1} p$

X has Geometric dist

$X \sim \text{Geometric}(p)$

Poisson

$X \sim \text{Poisson}(\lambda)$ $\lambda > 0$ if $f(x) = P(X=x) = \dfrac{e^{-\lambda}\lambda^x}{x!}$ $x = 0, 1, \ldots$

Poisson models eg. # of radioactive particles emitted by radioactive source in minute where λ is average emissions per minute.

Random Variables

Defn A Random Variable (rv) is a function, X,

$$X : \Omega \to \mathbb{R}$$

Ω: sample space; $\mathbb{R}$: reals

Informally: any observed random quantity

Defn

If r.v. X has finite outcomes $\{x_1 \cdots x_n\}$ or countably many outcomes it is discrete.

Write $f(x) = P(X=x)$ or $f_X(x) = P(X=x)$ for probabilities of outcomes. $f(x)$ is probability mass function

Ex Flip a coin n times and let X be # H's.

Suppose coin is biased s.t. $P(H) = p$; $P(T) = q = 1-p$.

Then $\Omega = \{H,T\} \times \{H,T\} \times \cdots \times \{H,T\}$ (n times)

$X(\underset{\omega \in \Omega}{HHTTHTH}) = 4$

Possible outcomes: $x = 0, 1, \cdots n$

$$f(x) = P(X=x) = \binom{n}{x} p^x q^{n-x} \quad x = 0 \cdots n$$

This is binomial dist. Write $X \sim \text{Binomial}(n, p)$ as shorthand.

Poisson cont

In ~~modelling~~ this situation divide interval (minute) into n pieces

and suppose each subinterval has emission with prob. $\frac{\lambda}{n}$

$\Rightarrow$ total # of S's (emissions) would be

$$X_n \sim \text{Binomial}\left(n, \frac{\lambda}{n}\right)$$

(Note "ave" # of S's is λ as it should be)

Can show $\lim_{n\to\infty} \text{Binomial}\left(n, \frac{\lambda}{n}\right) = \text{Poisson}(\lambda)$

$$\lim_{n\to\infty} P(X_n = x) = \lim_{n\to\infty} \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \overset{\text{A/H}}{=} \frac{e^{-\lambda}\lambda^x}{x!} \quad x = 0, 1, \ldots$$

Continuous Random Variables

X is continuous if it takes values

For any discrete d.s. must have $\sum_x f(x) = 1$

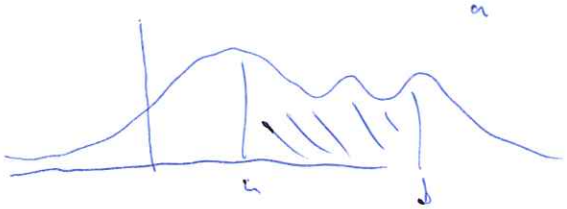
Continuous Random Variables

Continuous r.v's have a continuum of possible outcomes.

Eg $[0,1]$, $[0,\infty)$, $\mathbb{R}$, etc.

A continuous r.v. X has prob. density function, $f(x)$, s.t.

$$P(a < X < b) = \int_a^b f(x)\,dx.$$



pdf f must have $\int_{-\infty}^{+\infty} f(x)\,dx = 1$

NB If X has pdf $f(x)$

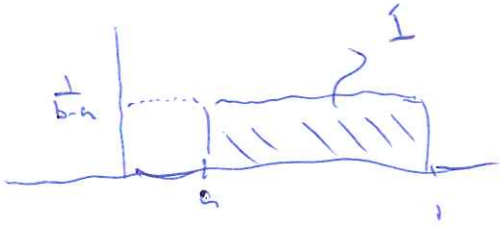
then $P(X = x) = \int_x^x f(y)\,dy = 0.$

Important Continuous Distributions

Uniform

If $X \in [a,b]$ and no outcome preferred over any other

$X \sim Unif(a,b)$ with pdf $f(x) = \begin{cases} \frac{1}{b-a} & x \in [a,b] \\ 0 & o.w. \end{cases}$



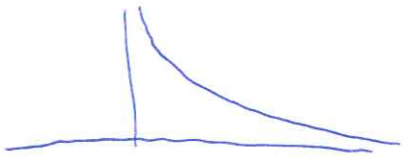
Exponential

Recall radioactive particles where λ was ave # emissions per unit time.

How long do we wait for 1st emission? Let X be time we wait.

$X \sim Exponential(\lambda)$ $f(x) = \lambda e^{-\lambda x}$ $x \geq 0$

0 $o.w.$



Normal

X has normal dist. w/ parameters μ, σ^2 if

$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$

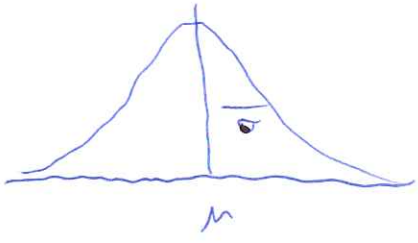
① Sums of indep. rv's approx Normal. (Central Limit Thm)

② Normals are easy to compute with

Eg $X_1 \sim N(\mu_1, \sigma_1^2)$ $X_2 \sim N(\mu_2, \sigma_2^2)$

X_1, X_2 indep.

$X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$



B555: Homework 1, Due Jan. 30 in class

From Wasserman work problems
Ch 1: 5,6,9,11,12,15,17,19,20
Ch 2: 7,11,14,15,21

B555 HW 1

Chapter 1

5) $\# \Omega = \{x_1 \cdots x_n : x_n = H,\ x_i = H \text{ some } i < n,\ x_j = T\ j \neq i, j \neq n\}$

Let X be the flip giving the 2nd H.

$P(X = x) = (x-1)\left(\frac{1}{2}\right)^x \quad x = 2, 3, \cdots$

6) The uniform dist must satisfy $f(x) = c$ for $x = 0, 1, \cdots$

If $c = 0$ then $\sum f(x) = 0$

If $c > 0$ then $\sum f(x) = \infty$

So can't satisfy $\sum_x f(x) = 1$.

9) (1) $P(A|B) = \frac{P(A,B)}{P(B)} > 0$

(2) $P(\Omega|B) = \frac{P(\Omega,B)}{P(B)} = \frac{P(B)}{P(B)} = 1$

(3) Suppose $A_1, A_2, \cdots$ are disjoint events. then

$$P\left(\bigcup_{i=1}^{\infty} A_i \mid B\right) = \frac{P\left(\left(\bigcup_{i=1}^{\infty} A_i\right) \cap B\right)}{P(B)} = \frac{P\left(\bigcup_{i=1}^{\infty} (A_i \cap B)\right)}{P(B)}$$

$$= \sum_{i=1}^{\infty} \frac{P(A_i \cap B)}{P(B)} = \sum_{i=1}^{\infty} P(A_i \mid B)$$

11) Let A, B be independent. Then since $A^c \cap B^c = (A \cup B)^c$

$$P(A^c \cap B^c) = 1 - [P(A) + P(B) - P(A)P(B)] = (1 - P(A))(1 - P(B))$$
$$= P(A^c)P(B^c)$$

12) Let $C \in \{1,2,3\}$ be chosen card and $S \in \{\text{Red}, \text{Green}\}$ be the color we see. $P(C = i) = 1/3 \quad i = 1,2,3$.

$$P(S = \text{Green} \mid C = i) = \begin{cases} 1 & i = 1 \\ 0 & i = 2 \\ 1/2 & i = 3 \end{cases}$$

$$P(C = i \mid S = \text{Green}) = \frac{P(C = i, S = \text{Green})}{P(S = \text{Green})} = \begin{cases} \frac{1/3}{1/2} & i = 1 \\ 0 & i = 2 \\ \frac{1/6}{1/2} & i = 3 \end{cases}$$

$$P(\text{other side Green} \mid S = \text{Green}) = P(C = 1 \mid S = \text{Green}) = 2/3.$$

15) a) Let N = # children with blue eyes $N \sim \text{Binomial}(3, 1/4)$

$$P(N \geq 2 \mid N \geq 1) = 1 - P(N < 2 \mid N \geq 1) = \frac{P(N = 1)}{P(N \geq 1)} = \frac{3\left(\frac{1}{4}\right)\left(\frac{3}{4}\right)^2}{1 - \left(\frac{3}{4}\right)^3}$$

b) The number of blue-eyed children from the remaining 2 children is $\text{Binomial}(2, 1/4)$.

$$P(\text{desired event}) = 1 - \left(\frac{3}{4}\right)^2$$

17)

Write $P_C(\cdot)$ for the prob. given C.

$$P(A,B,C) = P(A,B|C)P(C) = P_C(A|B)P_C(B)P(C)$$

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

19) $P(W) = .5$; $P(M) = .3$; $P(L) = .2$

$P(V|W) = .82$; $P(V|M) = .65$; $P(V|L) = .5$

$$P(W|V) = \frac{P(W)P(V|W)}{P(W)P(V|W) + P(M)P(V|M) + P(L)P(V|L)}$$

$$= \frac{.41}{.41 + .195 + .1} = \frac{.41}{.705}$$

20) a) $P(H) = \sum_{i=1}^{5} P(C_i)P(H|C_i) = \frac{1}{5}\left(0 + \frac{1}{4} + \frac{1}{2} + \frac{3}{4} + 1\right) = 1/2$

$$P(C_i|H) = \frac{P(C_i)P(H|C_i)}{P(H)} = \begin{cases} 0 & i=1 \\ \frac{1}{10} & i=2 \\ \frac{1}{5} & i=3 \\ \frac{3}{10} & i=4 \\ \frac{4}{10} & i=5 \end{cases}$$

b) $P(H_2|H_1) = \cancel{P(H_2|C_i)P(C_i|H_1)}$ $0 \cdot 0 + \frac{1}{10}\frac{1}{4} + \frac{2}{10}\frac{1}{2} + \frac{3}{10}\frac{3}{4} + \frac{4}{10}1 = 3/4$

c)

$$P(C_i|B_4) = \frac{P(C_i, B_4)}{P(B_4)} \qquad P(C_i|B_4) = \begin{cases} 0 & i=1 \\ \frac{\frac{1}{5}\left(\frac{3}{4}\right)^3\frac{1}{4}}{41/80} & i=2 \\ \frac{\frac{1}{5}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)}{41/80} & i=3 \\ \vdots \end{cases}$$

$$P(B_4) = \frac{1}{5}0 + \frac{1}{5}\left(\frac{3}{4}\right)^3\frac{1}{4} + \frac{1}{5}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right) + \frac{1}{5}\left(\frac{1}{4}\right)^3\frac{3}{4} + \frac{1}{5}0 = \frac{41}{80}$$

7) $P(Z > z) = P(X > z, Y > z) = (1-z)^2$

$$f_Z(z) = \frac{d\,P(Z \le z)}{dz} = \frac{d\,1-(1-z)^2}{dz} = 2(1-z) \quad 0 < z < 1$$

11) c) $P(X=0, Y=0) = 0$ yet $P(X=0) = P(Y=0) = 1-p$.

b) Note that $X+Y = N$ so

$$P(X=x, Y=y) = P(X=x, Y=y, N=x+y) = P(X=x \mid N=x+y)\,P(N=x+y)$$

$$= \frac{e^{-\lambda}\lambda^{x+y}}{(x+y)!}\binom{x+y}{x} p^x (1-p)^y$$

$$= \frac{e^{-\lambda}\lambda^x p^x}{x!}\,\frac{(1-p)^y}{y!}$$

The joint pmf factors into $g(x)h(y)$ so must be independent.

However, can show more directly by computing marginals.

$$P(X=x) = \sum_{n=0}^{\infty} P(X=x \mid N=n)\,P(N=n) = \sum_{n=0}^{\infty}\binom{n}{x} p^x (1-p)^{n-x}\frac{e^{-\lambda}\lambda^n}{n!}$$

$$= \sum_{n=x}^{\infty} \frac{p^x (1-p)^{n-x} e^{-\lambda}\lambda^n}{x!\,(n-x)!}$$

$$= \sum_{n=0}^{\infty} \frac{p^x (1-p)^n e^{-\lambda}\lambda^{x+n}}{x!\,n!}$$

$$= \frac{e^{-p\lambda}(p\lambda)^x}{x!} \sum_{n=0}^{\infty} \frac{e^{-(1-p)\lambda}((1-p)\lambda)^n}{n!} \qquad \sum_{n=0}^{\infty} \frac{e^{-(1-p)\lambda}((1-p)\lambda)^n}{n!} = 1$$

$$= \frac{e^{-p\lambda}(p\lambda)^x}{x!}$$

$X \sim \text{Poisson}(p\lambda)$ and by symmetry $Y \sim \text{Poisson}((1-p)\lambda)$

Indep follows by direct comparison of $P(X=x, Y=y)$ and $P(X=x)P(Y=y)$

Chapter 2 cont

14) $F_R(r) = P(R \le r) = r^2 \implies f_R(r) = 2r \quad 0 < r < 1$

15) $Y = F(X)$. note $0 \le Y < 1$

$P(Y \le y) = P(F(X) \le y) = P(X \le F^{-1}(y)) = y$

(By defn $F^{-1}(y)$ is # s.t. $P(X \le F^{-1}(y)) = y$)

$\implies Y = F(X) \sim Unif(0,1)$

Calculation shows if $X \sim F$ then $F(X) \sim Unif(0,1)$

Reverse calc shows ~~$U \sim Unif(0,1)$~~ $\implies F^{-1}(U) \sim F$

For $Exp(\beta)$ dist.

$f(x) = \frac{1}{\beta} e^{-x/\beta} \implies F(x) = 1 - e^{-x/\beta}$

$\implies F^{-1}(y) = -\beta \log(1-y)$

So if $U \sim Unif(0,1)$ then $-\beta \log(1-U) \sim Exp(\beta)$

21) Let $Y = \max(X_1 \cdots X_n)$ where $X_1 \cdots X_n \overset{iid}{\sim} Exp(\beta)$

$F_Y(y) = P(Y \le y) = P(X_1 \le y, \cdots X_n \le y) = (1 - e^{-y/\beta})^n$

$\implies f_Y(y) = \frac{n}{\beta}(1 - e^{-y/\beta})^{n-1} e^{-y/\beta} \quad y > 0$

Mean + Variance

For discrete r.v. X with pmf $f(x)$ the expectation of X is

$$E(X) = \mu = \sum_x x f(x)$$

IF X is continuous with pdf $f(x)$ " " " " "

$$E(X) = \mu = \int_{-\infty}^{+\infty} x f(x)\,dx$$

$E(X) = \mu$ is the average value or "mean" of dist.

For discrete X with pmf $f(x)$ the variance of X is

$$V(X) = \sigma^2 = \sum_x (x-\mu)^2 f(x) = \text{ave sq dist. from } \mu$$

so $V(x)$ measures spread,

For continuous r.v. X with pdf $f(x)$

$$V(X) = \sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)\,dx$$

Back to Normal

Recall $X \sim N(\mu, \sigma^2)$ if $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2\sigma^2}(x-\mu)^2}$

The "parameters" of normal, μ, σ^2, are in fact mean + variance

$$\int_{-\infty}^{+\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\,dx = \mu$$

$$\int_{-\infty}^{+\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\,dx = \sigma^2$$

Computational Ex

As stated before sums of ~~[illegible]~~ several indep. r.v's are approx normal.

~~[illegible]~~ More precisely: **CLT**

$X, X_1, \dots$ sequence of indep. rv's with same dist. with mean μ + var σ^2. Then for any $[a,b]$

$$P\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \in [a,b] \right) \xrightarrow{n\to\infty} P(Z \in [a,b])$$

where $Z \sim N(0,1)$

Ex

$X_1, \dots X_5 \overset{\text{indep.}}{\sim} \text{Unif}(0,1)$

$Y = \sum_{i=1}^{5} X_i$

① $E(X_i) = \int_0^1 x \cdot 1 \, dx = 1/2$

② $V(X_i) = \int_0^1 (x - \frac{1}{2})^2 \cdot 1 \, dx = 1/12$

③ $E(Y) = E\left(\sum_{i=1}^{5} X_i\right) = \sum_{i=1}^{5} E(X_i) = 5/2$

④ $V(Y) = V\left(\sum_{i=1}^{5} X_i\right) = \sum_{i=1}^{5} V(X_i) = 5/12$

⑤ $Y \overset{\text{approx}}{\sim} N\left(5/2, 5/12\right)$

Let's simulate Y many times and compare

a) $\hat{P}(Y \le y) = \dfrac{\#\text{ trials} \le y}{n}$

with

b) $P(N(5/2, 5/12) \le y)$ [can compute this]

An Introduction to R

To get R:

1. Download R (it's free) from the website http://cran r-project org There are versions for Linux, Windows and Mac.
2. R will be installed in the public computing cluster in room 109 in the Informatics building.

R can be used as a calculator:

Try typing the following expressions at the command line (followed by return): (> is the command prompt).

```
> 5+3
> 10*10
> log(9.4)
> exp(exp(exp(20)))  # R is only human! (anything following '#' is a comment)
```

R has most any mathematical function you can think of such as sqrt(), sin() ... mostly with easily guessable names. Expressions using the logical operators ==, !=, <, > give Boolean values (T,F)

```
> 4 > 3       # this evaluates to T (true)
> 1 == exp(0) # so does this
> 1 != exp(0) # this evaluates to F (false)
```

It is possible to have variables that hold values in your program. Most strings beginning with an alphabet character will be treated as variables. The assignment operator is <- (a "less than" followed by "minus"). Try typing the following lines in succession

```
> x <- 3
> y <- x*x+x
> y    # print the value of y
```

Vectors

One of the nicest aspects of R is the way it handles vectors. Here are a several ways to create vectors:

```
> x <- 1:100                  # x is now the vector (1,2,...,100)
> y <- seq(-pi,pi,length=100) # y consists of 100 evenly spaced values from -pi to pi
> z <- c(1,4,8,20)            # z is the vector (1,4,8,20)
> a <- x+y                    # vectors of same length can be added, multiplied, etc.
> b <- 4*x                    # this is interpreted correctly too
```

Random Number Generation

R has lots of built-in functions for doing things with random numbers. For instance

```
> x <- runif(100)   # creates a vector of 100 (uniformly distributed) random numbers between 0 and 1.
> punif(v)          # is the probability that a Unif(0,1) rand number is less than v
> qunif(u)          # gives the uth quantile of a Unif(0,1).  More on this later.
```

There are similar functions for a variety of other distributions including the normal(0,1) (rnorm,pnorm,qnorm) Cauchy (rcauchy, pcauchy, qcauchy), Exponential, Binomial, Poisson, and others.

Subsets

```
> x <- runif(100)   # creates a vector of 100 Unif(0,1) random numbers
> x[1]              # the first element of x
> x[c(1,3,5)]       # a vector containing 1st, 3rd and 5th elements of x
> y <- x > .5       # a 100-long vector of Boolean values y[i] is T iff x[i] > .5
> z <- x[x>.5]      # the ''x's'' that are greater than 5
```

Plotting Try the following

```
> x <- seq(0,1,length=100)
> y <- x^2                    # y = x squared
> plot(x,y)                   # plot with (x[1],y[1]) \ldots, (x[100],y[100])
> plot(y,x)
> plot(y)      # same as plot(1:length(y),y)
```

Source Files You will want to write simple programs in R and this always requires some trial, error and iteration. I recommend the following procedure: Create a "source" file in any text editor containing your R commands. This could be emacs or the Windows "Notepad" or whatever you are comfortable using. Suppose you create the following file named "myprog.R"in your editor:

```
len <- 100
x <- runif(len,-.5,.4)
y <- cumsum(x)   # y[1] = x[1], y[2] = x[1]+x[2], etc.
plot(exp(y))
title("my stock price")
print("history is: ")
print(y)
```

This technique allows you to write a program in the usual incremental way. If you want to get a hard copy of the printout and the plot (for example, to submit as your homework), do the following

```
> postscript("myplot.ps")   # write plot in the postscript file ''myplot.ps''
> sink("myout.txt")         # write text output to ''myout.txt''
> source("myprog.R")        # run the program you created
> dev.off()                 # redirect plots to screen.  Don't forget this!
> sink()                    # redirect output to screen. ditto.
```

A fun example Suppose two decks of cards are shuffled. The first deck is arranged in a line, face up, and the second deck is arranged in a similar line right below. Count the number of places where the two decks have the same card. What is the probability that there are no matches between the two decks? (This is a hard calculation to do). One way to estimate this probability would be to perform the experiment many times and observe the proportion of times the event occurs.

```
> trials <- 1000          # number of trials
> zeros <- 0              # counts the number of times the ''no match'' event occurs
> for (i in 1:trials) {   # all statements in the ''for'' loop are executed ---
                          # once for each value of i in 1:trials
>   x <- 1:52;
>   deck1 <- sample(x,52,replace=F)  # a random permutation of the ''cards''
>   deck2 <- sample(x,52,replace=F)  # another random permutation
>   matches <- sum(deck1==deck2)     # number of matches
>   if (matches == 0) zeros <- zeros+1  # if (matches == 0) we had a no matches  so count it
> }
> print("my estimate is:")
> print(zeros/trials);    # the proportion of times the ''no match'' event occured
```

Quitting and help

```
> help("rnorm")  # gives information about the function rnorm.  Of couse this works
>                # for other functions too.
> q()  # quitting the program. Hope you had fun.
```

Joint Distributions

Several rvs $X_1 \cdots X_n$, perhaps interdependent, have a joint distribution, $f(x_1 \cdots x_n)$

For $X_1 \cdots X_n$ discrete $f(x_1 \cdots x_n)$ is the joint pmf defined by

$$f(x_1 \cdots x_n) = p(X_1 = x_1, X_2 = x_2, \cdots X_n = x_n)$$

For $X_1 \cdots X_n$ continuous $f(x_1 \cdots x_n)$ is the joint pdf s.t.

$$p(a_1 < X_1 < b_1, \cdots, a_n < X_n < b_n) = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} f(x_1 \cdots x_n)\, dx_n \cdots dx_1$$

Must have $\sum_{x_1 \cdots x_n} f(x_1 \cdots x_n) = 1$ for discrete case

$$\int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f(x_1 \cdots x_n)\, dx_n \cdots dx_1 \quad \text{for continuous case.}$$

Marginal Distributions

For bivariate discrete X, Y can represent pmf as table

y \ x	0	1	
0	.1	.2	.3
1	.3	.4	.7
	.4	.6	

Summing cols/rows give "marginal" distributions (distributions of X or Y with no knowledge of other vble.)

More generally, for discrete $X \cdots X_n$ with pmf $f(x_1 \cdots x_n)$ the marginal distributions $f_{X_1}(x_1) \cdots f_{X_n}(x_n)$ are

$$f_{X_i}(x_i) = \sum_{x_1 \cdots x_{i-1}, x_{i+1} \cdots x_n} f(x_1 \cdots x_n).$$

For continuous $X_1 \cdots X_n$ with pdf $f(x_1 \cdots x_n)$ the marginal densities, $f_{X_1}(x_1) \cdots f_{X_n}(x_n)$ are

$$f_{X_i}(x_i) = \int_{x_1} \cdots \int_{x_{i-1}} \int_{x_{i+1}} \cdots \int_{x_n} f(x_1 \cdots x_n)\, dx_n \cdots dx_{i+1}\, dx_{i-1} \cdots dx_1$$

Independence

Random variables $X_1 \cdots X_n$ are independent if their joint is product of marginals.

$$f(x_1 \cdots x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$$

This holds for both continuous + discrete rv's.

Equivalently $X_1 \cdots X_n$ indep. if the events they generate are independent. That is $\{X_1 \in A_1\}, \{X_2 \in A_2\} \cdots \{X_n \in A_n\}$ are independent.

Important Examples

Multinomial

Discrete $X_1 \cdots X_n$ have a multinomial dist.

$X_1 \cdots X_n \sim \text{Multinomial}(p_1 \cdots p_n, n)$

Have n independent trials with K possible outcomes w/ probs $p_1 \cdots p_K$. Let X_k = # times outcome k occurs.

$X_1 \cdots X_K$ said to have multinomial dist $X_1 \cdots X_K \sim \text{Multinomial}(p_1 \cdots p_K, n)$

$$f(x_1 \cdots x_K) = P(X_1 = x_1 \cdots X_K = x_K) = \frac{n!}{x_1! \cdots x_K!} p_1^{x_1} p_2^{x_2} \cdots p_K^{x_K} \quad x_1 \cdots x_K = 0, 1, \cdots n \text{ s.t. } \sum x_K = n$$

Q: What is $f_{X_1}(x_1)$? A: Each trial either type 1 (p_1) or not ($1-p_1$)

$$\Rightarrow X_1 \sim \text{Binom}(n, p_1) \qquad f_{X_1}(x_1) = \binom{n}{x_1} p_1^{x_1} (1-p_1)^{n-x_1} \quad x_1 = 0 \cdots n$$

[$\sum_{\substack{x_2 \cdots x_K \\ \sum x_i = n}} \frac{n!}{x_1! \cdots x_K!} p_1^{x_1} \cdots p_K^{x_K} = \frac{n!}{x_1!(n-x_1)!} p_1^{x_1} (1-p_1)^{n-x_1}$]

Expectation + Covariance

For the $X_1 \cdots X_n$

Notation Often better to use vector notation and write

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$$ for collection of random variables.

Expectation + Covariance

For random vector $X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$ the expectation of X, $E(X) = \begin{pmatrix} EX_1 \\ \vdots \\ EX_n \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}$

where EX_i is expectation of ith vble.

That is

$$EX_i = \sum_{x_i} x_i f_{X_i}(x_i) = \sum_{x_1 \cdots x_n} x_i F(x_1 \cdots x_n) \quad (*\text{discrete})$$

$$EX_i = \int_{x_i} x_i f_{X_i}(x_i)\, dx_i = \int_{x_1} \cdots \int_{x_n} x_i F(x_1 \cdots x_n)\, dx_n \cdots dx_1 \quad (\text{continuous})$$

Covariance ① For X, Y rvs $Cov(X,Y) = E(X-\mu_X)(Y-\mu_Y)$ [the covariance] ③ Do correlation

For random vectors variance generalizes to covariance matrix.

④ For random vector $X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$ the covariance matrix, $\sum_{n \times n}$, is given by

$$\Sigma_{ij} = E(X_i - \mu_i)(X_j - \mu_j)$$

Observe that $\Sigma_{ii} = E(X_i - \mu_i)^2 = V(X_i)$ so diagonal elements of Σ are variances.

② Intuitively, if X_i, X_j vary together (one "high" (low) when other high (low)) then $(X_i - \mu_i)$ and $(X_j - \mu_j)$ tend to have same sign, so

$$\Sigma_{ij} = Cov(X_i, X_j) > 0$$

Alternatively, if X_i, X_j vary inversely (one high when other low) $(X_i - \mu_i)$ and $(X_j - \mu_j)$ tend to have opposite sign so

$$\Sigma_{ij} = Cov(X_i, X_j) < 0$$

⑤ Often write $X = \begin{pmatrix} X_1 \\ X_n \end{pmatrix}$ $\mu = \begin{pmatrix} \mu_1 \\ \mu_n \end{pmatrix}$

$\Sigma = Cov(X) = E(X-\mu)(X-\mu)^t = E$

$\begin{pmatrix} (X_1-\mu_1)(X_1-\mu_1) & (X_1-\mu_1)(X_2-\mu_2) & \cdots & (X_1-\mu_1)(X_n-\mu_n) \\ (X_2-\mu_2)(X_1-\mu_1) & \ddots & & \\ \vdots & & & \\ (X_n-\mu_n)(X_1-\mu_1) & & \cdots & (X_n-\mu_n)(X_n-\mu_n) \end{pmatrix}$

11

Important Facts about Expectations

① Expectation is linear. That is, "Expectation of sum is sum of expectations"

HW If $X_1 \cdots X_n$ have $EX_1 = \mu_1 \cdots EX_n = \mu_n$

$$E(\lambda_1 X_1 + \lambda_2 X_2 + \cdots + \lambda_n X_n) = \lambda_1 \mu_1 + \lambda_2 \mu_2 + \cdots \lambda_n \mu_n.$$

In vector notation, if $\lambda^t = (\lambda_1 \cdots \lambda_n)$ and $X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$

$$E(\lambda^t X) = \lambda^t E(X)$$

or, if $\underset{k \times n}{A}$ is matrix $\begin{pmatrix} a_{11} \cdots a_{1n} \\ a_{21} \cdots a_{2n} \\ \vdots \\ a_{k1} \cdots a_{kn} \end{pmatrix} = A$

$$E(AX) = A E(X)$$

② If X_1, X_2 independent, $E(X_1 X_2) = E(X_1) E(X_2)$

HW

Joint Normal

Z is standard normal if $Z \sim N(0,1)$. $\left[f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} z^2} \right]$

Suppose $Z_1 \cdots Z_n$ are standard normal + independent. If $Z_1 \cdots Z_n \overset{iid}{\sim} N(0,1)$ = "independent identically distributed"

$$Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix} \quad E(Z) = 0 = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \quad \text{and}$$

$$\mathrm{Cov}(Z_i, Z_j) = E Z_i Z_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \quad (E Z_i Z_j = E Z_i \, E Z_j)$$

So the covariance matrix of Z is $\underset{n \times n}{I}$ (identity matrix)

$$f(z_1 \cdots z_n) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}(z_1^2 + z_2^2 + \cdots z_n^2)} = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2} z^t z} = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\|z\|^2}$$

Correlation

Recall $|\sum x_i y_i| \leq \left(\sum x_i^2\right)^{1/2}\left(\sum y_i^2\right)^{1/2}$ Cauchy-Schwartz

Also by C.S.

$$|Cov(X,Y)| = \left|\sum_{x,y}(x-\mu_x)(y-\mu_y)f(x,y)\right| \leq \left(\sum_{x,y}(x-\mu_x)^2 f(x,y)\right)^{1/2}\left(\sum_{x,y}(y-\mu_y)^2 f(x,y)\right)^{1/2}$$

$$= Var(X)^{1/2}\, Var(Y)^{1/2} = \sigma_X \sigma_Y$$

$\Rightarrow$

$$-1 \leq \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \leq 1$$

Define correlation of X, Y to be

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \rho_{XY}$$

Have $-1 \leq \rho_{XY} \leq 1$. If $\rho_{XY} = 0$ X, Y said to be uncorrelated

Correlation + Independence

If X, Y indep. $Cov(X,Y) = E(X-\mu_X)(Y-\mu_Y) = E(X-\mu_X)\,E(Y-\mu_Y) = 0$

Indep $\Rightarrow$ Uncorrelated

~~$\Leftarrow$~~

HW shows examp.

Multivariate Normal

$X = \begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix}$ has multivariate normal dist. $\wedge$ $X \sim N(\mu, \Sigma)$ w mean μ and cov Σ if

$$f(x) = f\begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix} = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^t \Sigma^{-1}(x-\mu)}$$

(1) By direct calc. can show $E(X) = \int x f(x)\,dx = \mu$

$$Cov(X) = \int (x-\mu)(x-\mu)^t f(x) = \Sigma$$

(2) Consider $Z_1 \cdots Z_k \overset{iid}{\sim} N(0,1)$ $Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_k \end{pmatrix}$ has

$$f(z) = \frac{1}{(2\pi)^{k/2}} e^{-\frac{1}{2}z^t z} \Rightarrow Z \sim N(0, I)$$

(3) For invertible $A_{k\times k}$ AZ has

a) $E(AZ) = AE(z) = A0 = 0$

b) $\Sigma =$
$Cov(AZ) = E\left(AZ(AZ)^t\right) = E\left(AZZ^tA^t\right) = AEZZ^tA^t$

$= AIA^t = AA^t$

$(\Sigma = AA^t \Rightarrow \Sigma^{-1} = A^{-t}A^{-1})$

(4) Let $X = AZ$ What is pdf of X?

$$f_X(x) = \frac{1}{|A|} f_Z(A^{-1}x) = \frac{1}{(2\pi)^{k/2}|A|} e^{-\frac{1}{2}x^tA^{-t}A^{-1}x} = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}x^t\Sigma^{-1}x}$$

$\Sigma = AA^t$ $\Rightarrow X \sim N(0, \Sigma = AA^t)$

$|\Sigma| = |A||A| \Rightarrow |\Sigma|^{1/2} = |A|$

(5) Any time have $X \sim \mathcal{N}(\mu, \Sigma)$ can think of $X = AZ + \mu$

where A is chosen s.t. $AA^t = \Sigma$.

Do pairs.r to demonstrate.

Conditional Distributions

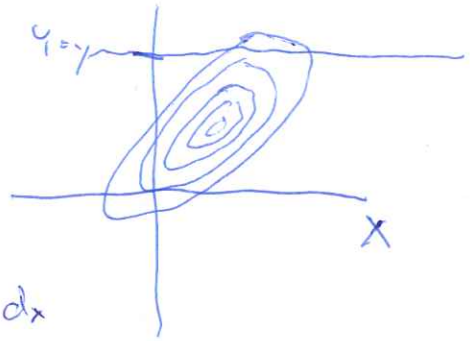
The conditional dist. of X given Y, $f(x|y) = f_{X|Y=y}(x|y)$ is pmf or pdf for X when known that $Y = y$.

For X, Y discrete

$$f(x|y) = \frac{P(X=x, Y=y)}{P(Y=y)} = P(X=x|Y=y) = \frac{f(x,y)}{f_Y(y)}$$

For X, Y continuous still have

$$f(x|y) = \frac{f(x,y)}{f_Y(y)}$$



where $P(X \in A | Y=y) = \int_{x \in A} f(x|y)\,dx$

Similarly, for X, Y, Z $f(x,y|z) = \frac{f(x,y,z)}{f_Z(z)}$

is conditional pmf, pdf of X, Y given $Z = z$.

Conditional Independence

For X, Y, Z, X and Y conditionally independent given $Z = z$ if X, Y indep in conditional dist $f(x,y|z)$. That is

$$f(x,y|z) = f(x|z)\,f(y|z).$$

Do simpson.r
Could have all departments giving equal acceptance rates to M/F, yet overall having different acceptance rates to M/F

X, Y indep

Z

X, Y cond indep given Z

NB

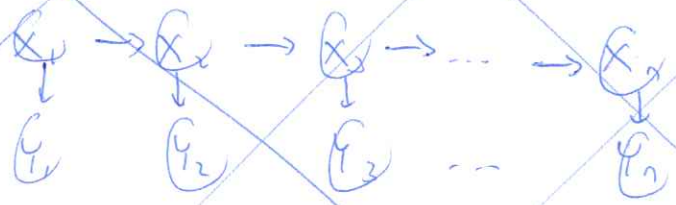
Important Graphical Models

(1) $X_1 \cdots X_n$ respect

$$p(x_1 \cdots x_n) = p(x_1)\, p(x_2|x_1)\, p(x_3|x_2) \cdots p(x_n|x_{n-1})$$

$X_1 \cdots X_n$ form a Markov Chain.

(2) $X_1 \cdots X_n, Y_1 \cdots Y_n$ respect



$$p(x_1 \cdots x_n y_1 \cdots y_n) = p(x_1)\, p(x_2|x_1) \cdots p(x_n|x_{n-1})\, p(y_1|x_1)\, p(y_2|x_2) \cdots p(y_n|x_n)$$

X, Y are a hidden Markov Model.

Mixture Distributions

Read Ch 3 Wasserman (not moment generating fns)

Have M different prob. models $p_1(x) \cdots p_M(x)$. To generate X:

choose model 1 with prob q_1

" " 2 " " q_2

⋮

" " M " " q_M $\quad \left(\sum_{m=1}^{M} q_m = 1\right)$

and sample from chosen model.

What is dist. for X obtained this way? By LoTP

$$p(x) = \sum_{m=1}^{M} p(m)\, p(x|m) = \sum_m q_m\, p_m(x)$$

$p(x)$ is mixture of $p_1 \cdots p_M$ with mixing weight $q_1 \cdots q_M$

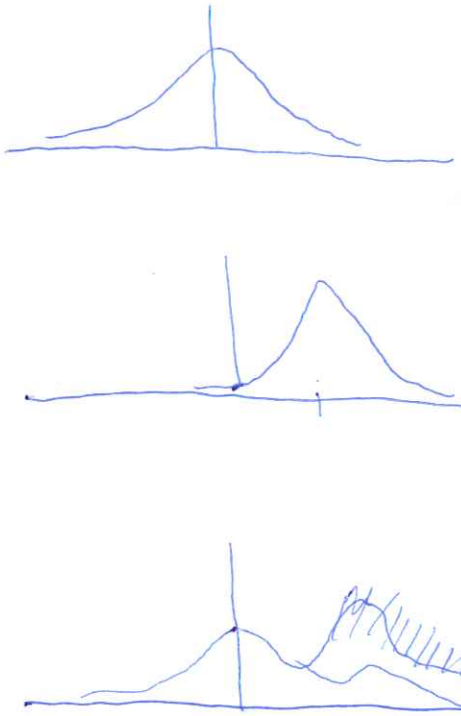
Ex

Have $p_1(x) = \mathcal{N}(x; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$

$p_2(x) = \mathcal{N}(x; 1, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-1)^2}$

Choose model 1 with prob 3/4

" " 2 " " 1/4



Many (maybe most) real world distributions are mixtures.

Population contains several types of individuals, each with characteristic dist.
The mixing weights are the proportions of each type.
In sampling characteristic from pop, choose type according to ~~k~~ dist
representation. Then sample from type dist.

① Ex

Find mixture dists. in NYT Game of Thrones

Likely Mixtures: Jeffrey Baratheon, Cyburn, Bran Stark, Jeor Mormont.

Are judgements of Goodness + Beauty independent?

② Ex Poisson_binom.r

$Z \sim \text{Poisson}(\lambda)$ X, Y = H's (T's) from Z flips

$P(H) = p$ $P(T) = q$ $X \sim \text{Binom}(Z, p)$ $Y = Z - X$

$$P(X=x) = \sum_{z=0}^{\infty} P(Z=z)\, P(X=x \mid Z=z)$$

Data, $= \sum_{z=0}^{\infty} \frac{e^{-\lambda}\lambda^z}{z!} \binom{z}{x} p^x (1-p)^{z-x}$

Observe X, Y indep.

Conditional Independence

Write $p(x)$, $p(y|x)$ etc. for pmf, pdf, conditional pmf/pdf etc.

(1) For rv's X, Y, Z, X, Y conditionally independent given Z if X, Y independent under conditional dist given $Z = z$. That is

$$p(x, y \mid z) = p(x \mid z)\, p(y \mid z)$$

(2) **Ex**

Suppose roll die (Z). Then take 2 fair coins and let

X = # H's in Z flips of coin 1

Y = # " " " " " " 2

When X is high Y will also tend to be high so (informally) X, Y not independent.

But once we know Z (conditioned on Z) X, Y independent.

(4) **Note**

If X, Y independent then conditioning on X adds no info on Y. That is

$$p(y \mid x) = p(y) \qquad \left[\; p(y|x)\,p(x) = p(x,y) = p(x)\,p(y) \;\right]$$

Similarly, if X, Y conditionally independent given Z, once we know Z additional knowledge of X gives no new info about Y. Ie

$$p(y \mid x, z) = p(y \mid z)$$
$$p(x \mid y, z) = p(x \mid z)$$

So factorization simplifies

$$p(x, y, z) = p(z)\, p(y|z)\, p(x|y,z) = p(z)\, p(y|z)\, p(x|z).$$

(6) **NB**

X, Y cond ind given Z $\not\Rightarrow$ X, Y indep

$\not\Leftarrow$

What is a reasonable C.I. relationship?

Given Y, X, Z indep.

(3) **Ex**

X = person owns ash tray

Y = person smokes

Z = person gets lung cancer

(5) Do Simpson's (Simpson's Paradox)

Bayesian Network

As shown in HW can always factor joint dist through "chain rule":

$$p(x,y,z) = p(x)\,p(y|x)\,p(z|y,x)$$

Factorization not unique since other possible orderings:

$$p(x,y,z) = p(z)\,p(y|z)\,p(x|y,z)$$

There are $n!$ possible orderings of n variables so $n!$ possible factorizations

But they are all equal. Eg

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

(Box)

Cond. Ind. makes factorization simpler.

For instance, if X_n cond. ind of $X_1 \cdots X_{n-2}$ given X_{n-1} then

$$p(x_n | x_1, x_2 \cdots x_{n-1}) = p(x_n | x_{n-1})$$

If the $\{X_i\}$ are binary r.v's.

LHS: 2^{n-1} Bernoulli distributions RHS: 2 Bernoulli dists.

Roughly speaking ^ LHS requires amount of data (training) req. to 2^{n-1}

RHS " " " " " " 2

LHS may well be intractable for large n.

Bayesian Network uses cond. indep. to simplify factorization.

Eg, suppose for X, Y, Z, W

Y cond ind of X given W $\quad p(y|x,w) = p(y|w)$

Z cond inde of W given X, Y $\quad p(z|w,x,y) = p(z|x,y)$

$p(w,x,y,z) = p(w)\,p(x|w)\,p(y|w,x)\,p(z|w,x,y)$ "

$= p(w)\,p(x|w)\,p(y|w)\,p(z|x,y)$ "

Why important? See Box

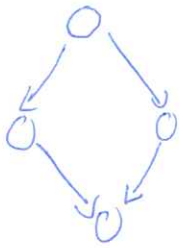
Graphical Representation

Factorization of joint dist can be represented by directed acyclic graph (DAG)

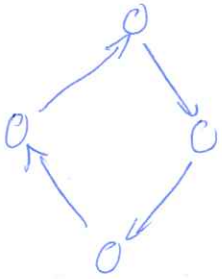
G: Graph is collection of nodes/vertices (random variables) and edges (dependency)

D: Edges are directed

A: no directed cycles



okay



not okay.

Ex

$$p(x,y,z) = p(x)\,p(y|x)\,p(z|x,y)$$

DAG Reg.

$$p(x,y,z) = p(x)\,p(y|x)\,p(z|y)$$

[Given Y, X, Z cond. indep]

Parents of variable in graph are variables we condition on.

More generally

If $A \subseteq \{1, \dots n\}$ write $X_A = \{X_i : i \in A\}$

still write X_i for $X_{\{i\}}$

Suppose $X_1 \cdots X_n$ r.v's ~~w/ dist. described by ...~~

The following statements are equivalent

(1) $X_1 \cdots X_n$ respects a given DAG

(2) $g(x_1 \cdots x_n) = \prod_{i=1}^{n} g(x_i \mid x_{P(i)})$ where $P(i)$ are "parents" of i in DAG (all nodes connected to i)

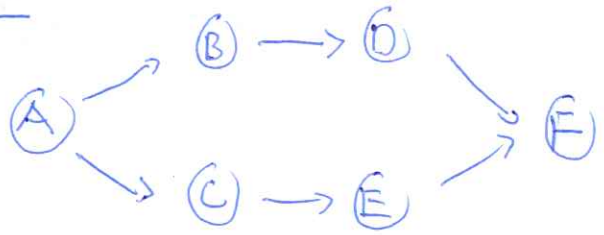
(3) Let $A(i)$ be the ancestors of i. That is

$j \in A(i)$ iff there is directed path from j to i

$j \to k_1 \to k_2 \to \cdots \to i$

Given $P(i)$ X_i cond indep. of $A(i) \setminus P(i)$ (set subtraction.)

Ex



$g(a,b,c,d,e,f)$ respects graph

$g(a,b,c,d,e,f) = g(a)\, g(b|a)\, g(c|a)\, g(d|b)\, g(e|c)\, g(f|d,e)$

C.I's $P(F)$ $A(F) \setminus P(F)$

For ex. Given $\{D, E\}$, F cond indep. of $\{A, B, C\}$

with similar statements for other vbles.

Simulation

Can think of $g(x_1 \cdots x_n) = \prod_{i=1}^{n} g(x_i \mid x_{P(i)})$ as algorithm for simulating from joint dist.

① For each node with no parents simulate from marginal dist.

Repeat: Find node i s.t. all of i's parents have been simulated.

Draw X_i from $g(x_i \mid x_{P(i)})$

B555: Homework 2

1. Suppose that X is a discrete n-dimensional random vector

$$X = \begin{pmatrix} X_1 \\ X_1 \\ \vdots \\ X_n \end{pmatrix}$$

and that $a \in \Re^n$ (a is a constant n-dimensional vector). Show $E(a^t X) = a^t E(X)$ where a^t denotes the transpose of a. (The result also holds for continuous random vectors).

2. Let X, Y be independent continuous random variables with density $f(x, y)$, where $f(x, y) = g(x)h(y)$ for some functions g, h. Show that X and Y are independent.

3. For this problem, recall that random variables X, Y are conditionally independent given Z if X and Y are independent under the conditional distribution $f(x, y|z)$ (the distribution of X, Y having observed $Z = z$.

   Suppose that we consider random minutes from randomly chosen drivers to determine three binary random variables:

$$\begin{aligned} H &= \text{the driver is within 1 mile of home} \\ M &= \text{the driver is on a main road} \\ A &= \text{the driver has had an accident in the sampled minute} \end{aligned}$$

   Suppose that $P(H = 1) = .8$ along with $P(M|H)$ and $P(A = 1|, H, M)$ in the following two tables.

$P(M\|H)$	$M = 0$	$M = 1$
$H = 0$	.1	.9
$H = 1$	.6	.4

$P(A = 1\|H, M)$	$M = 0$	$M = 1$
$H = 0$	.0003	.0008
$H = 1$	.0003	.0008

   (a) Compute $P(A = 1|H = 0)$ and $P(A = 1|H = 1)$ Are accidents more likely near home?

   (b) Compute $P(A = 1|M = 0)$ and $P(A = 1|M = 1)$. What conditional independence relation can you conclude?

   (c) Give a cause and effect hypothesis that is consistent with your conditional indepedence statement.

   (d) Suppose that a driver is driving on a stretch of highway (main road) on her way home. As she crosses the 1-mile-to-home road marker does she encounter an elevated probability of accident?

4. Simulation of Binomial

   (a) A random variable has $X \sim \text{Unif}(0, 1)$ when its density function is

$$f(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

   The R command "runif(n)" creates a vector of n independent Unif$(0, 1)$ outcomes. Use this function to simulate a sequence of 10,000 independent Binomial$(500, 1/2)$ random variables and plot the resulting histogram. You can plot a histogram of a data vector x with "hist(x)".

   (b) Letting your sequence of Binomial variables be $X_1, \ldots, X_n$, "standardize" these variables by taking $Y_n = (X_n - \mu)/\sigma$ where μ and σ are the mean and standard deviation of the variables. What are the mean and variance of the Y_n variables?

(c) Now simulate 10,000 *standardized* Binomials and plot the empirical cumulative distribution function $\hat{F}(x)$ where

$$\hat{F}(x) = \hat{P}(X \leq x) = \frac{|\{i : X_i \leq x\}|}{n}$$

Use "plot(x,type='l')" to plot with a solid line

(d) On the same plot show the cumulative distribution function of the standard normal. "pnorm(x)" gives the probability that a $N(0,1)$ variable is less than x. You can add to an existing plot with "lines(x)." Your two plots should look very similar. If X is a Binomial with large n, the fact that

$$\frac{X - np}{\sqrt{np(1-p)}} \overset{\text{approx}}{\sim} N(0,1)$$

is known as the normal approximation to the Binomial.

5. It is well-known that independent random variables are also uncorrelated.

6. Consider the variables $X = (X_1, X_2)$ having probability density function

$$p(x) = \begin{cases} \frac{1}{\pi} & x_1^2 + x_2^2 < 1 \\ 0 & \text{otherwise} \end{cases}$$

Show that X_1 and X_2 are uncorrelated but not independent.

7. Suppose that $X = (X_1, X_2) \sim N(\mu, \Sigma)$ where

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$$

(a) Show that in this case X_1, X_2 uncorrelated implies they are also independent. It might help to know that the marginal distribution of X_1 is normal with mean μ_1 and variance $\sigma_1^2 = \sigma_{11}$ with a simliar statement for X_2.

(b) Suppose, $X = (X_1, X_2, \ldots, X_p) \sim N(\mu, \Sigma)$. Show that if $\Sigma_{ij}^{-1} = 0$ then X_i and X_j are conditionally independent given *all* the other variables. (One can also show the stated conditional independence assumption implies $\Sigma_{ij}^{-1} = 0$ for when dealing with normal random variables).

(c) Suppose $X_1, X_2, \ldots, X_n$ defined by $X_1 = Z_1$ and $X_i = X_{i-1} + Z_i$ for $i = 2, \ldots, n$ where $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} N(0,1)$. This is known as a random walk.

i. Show that $(X_1, \ldots, X_n) \sim N(0, \Sigma)$ for some covariance matrix Σ.

ii. Without computing argue which elements of Σ^{-1} are 0.

8. Suppose that $X \sim \text{Unif}(0,1)$.

(a) Show that the cdf of $Y = \frac{-\log(X)}{\lambda} = 1 - e^{-\lambda y}$ and conclude that $Y \sim \text{Exponential}(\lambda)$.

(b) Suppose that $X_1, X_2 \ldots \overset{\text{iid}}{\sim} \text{Exponential}(\lambda)$. Interpret X_1 as the time we need to wait for the first event, X_2 as the *additional* time we wait for the 2nd event, X_3 as the additional time for the 3rd event and so on. In R take $\lambda = 3$ and simulate 100,000 times the number of events that occur in the interval $[0,1]$. Compute the empirical pmf, $\hat{p}(n)$ for $n = 0, 1, \ldots$ by

$$\hat{p}(n) = \frac{\#\text{ times experiments gives } n}{100000}$$

On the same plot compare these probabilities with the Poisson(λ) distribution. The process $N(t)$ that counts the number of events that occur before time t is known as a Poisson process.

9. Let M be a discrete variable with probability mass function $p(m)$. Suppose X is a continuous random vector that *depends* on M. Specifically, assume $X|M = m$ is $N(\mu_m, \Sigma_m)$.

(a) Even though M is discrete and X is continuous we can still get the joint distribution of M, X as $p(m, x) = p(m)p(x|m)$. In this problem we'll use the term density for dicrete, continuous and mixed cases. Compute

i. The joint density $p(m, x)$

ii. The marginal density $p(x)$.

iii. The conditional density $p(m|x)$

(b) Suppose that a population is composed of C classes $1, \ldots, C$. let $p(i)$ be the proportion of the population in class i so $p(1) + \ldots + p(C) = 1$. Suppose we take several measurments, X, on the population and that these measurements behave differently under the classes. We model the "class conditional" distribution of X as multivariate normal with mean μ_i and covariance Σ_i.

i. If we observe $X = x$ give the conditional probabilities $P(M = m|X = x)$.

ii. Suppose that you observe $X = x$ and want to choose the value of m, $\hat{M}$, that has the best chance of being right. What should $\hat{M}$ be?

Do learning_belief_network.

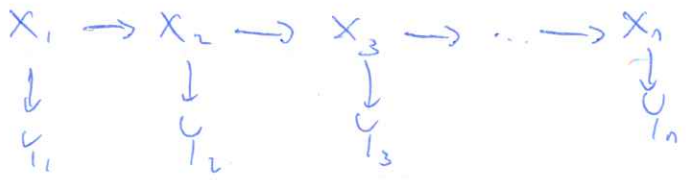
Important Graphical Models

① $X_1 \cdots X_n$ respect : $X_1 \to X_2 \to X_3 \to \cdots \to X_n$

$$p(x_1 \cdots x_n) = p(x_1)\, p(x_2|x_1)\, p(x_3|x_2) \cdots p(x_n|x_{n-1})$$

$X_1 \cdots X_n$ form Markov Chain

② $X_1 \cdots X_n$, $Y_1 \cdots Y_n$ respect



$$p(x_1 \cdots x_n, y_1 \cdots y_n) = p(x_1)\, p(x_2|x_1) \cdots p(x_n|x_{n-1})\, p(y_1|x_1) \cdots p(y_n|x_n)$$

X, Y are hidden Markov Model.

Classification

Have set of possible classes $1 \dots K$

Eg: Repub/Dem/Indep or Spam/Not Spam or ...

Have data vector, X, of observations whose distrib. depends on class

Observe $X = x$ and try to predict correct class.

Modeling

Let $c \in \{1, \dots k\}$ be class, modeled as r.v. with dist $p(c)$.

Let $p(x|c) = p_c(x)$ be class conditional distribution

(dist on X when class, c, known)

$p(c)$ is "prior" dist. on class [prior to seeing X)

$p(c|x)$ is "posterior" " " " [after " " "]

By Bayes' Rule

$$p(c|x) = \frac{p(c)\,p(x|c)}{p(x)} = \frac{p(c)\,p(x|c)}{\sum_{c'} p(c')\,p(x|c')} \quad \text{(LoTP)}$$

Note: $p(x) = \sum_c p(c)\,p(x|c)$ is mixture dist. with mixing weights $p(c)$

Optimal Classification (Bayes' Classifier)

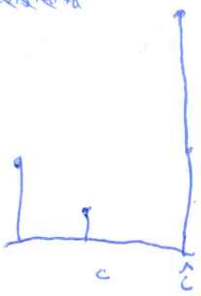
Notation: Will write $\hat{c}$ for our estimate of true class c.

Suppose we take

$$\hat{c} = \arg\max_c \eta(c|x) = \arg\max_c \frac{\eta(c)\eta(x|c)}{\eta(x)} = \arg\max_c \eta(c)\eta(x|c)$$

$\eta(x)$: constant

the c that maximizes $\eta(c|x)$



Fix x and consider errors

$$\eta(\text{error}|x) = \sum_{c \neq \hat{c}} \eta(c|x)$$

By choosing $\hat{c}$ as most likely class we minimize sum

$\Rightarrow$ $\hat{c} = \arg\max_c \eta(c|x)$ is optimal classifier.

Do: gaussian-classifier.r

coin_flip.class.error.r

Comment: Hard to implement Bayes Class.

Parameter Estimation

Bayes Classifier assumes we know class-conditional dists ($p_c(x) = \eta(x|c)$) though we don't. How to get these?

Maximum Likelihood Estimation (MLE)

Have family of distributions indexed by unknown parameter Θ.

Ex: $N(\Theta, \sigma^2)$, Exponential(Θ), Binomial(n, Θ), Poisson(Θ)

$N(\Theta_1, \Theta_2)$, ...

Observe random sample $x_1 \cdots x_n \overset{iid}{\sim} p_\Theta(x)$. MLE, $\hat{\Theta}_{MLE} = \hat{\Theta}_{MLE}(x_1 \cdots x_n)$

↳ Chooses value of Θ that best explains data $x_1 \cdots x_n$. That is $\hat{\Theta} = \arg\max_\Theta p_\Theta(x_1 \cdots x_n) = \arg\max_\Theta \prod_{i=1}^n p_\Theta(x_i)$.

Ex (Poisson)

Suppose $x_1 \cdots x_n \overset{i.i.d}{\sim} \text{Poisson}(\lambda)$. ✱

Fact

$X \sim \text{Poisson}(\lambda) \Rightarrow EX = \lambda$ $\quad \left[E(X) = \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!} = \sum_{x=1}^{\infty} \frac{e^{-\lambda}\lambda^x}{(x-1)!} \right.$

$\left. = \lambda \left(\sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} \right) = \lambda 1 = \lambda \right]$

$$\hat{\lambda}_{MLE}(x_1 \cdots x_n) = \arg\max_{\lambda} p_\lambda(x_1 \cdots x_n)$$

$$= \arg\max_{\lambda} \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$$ ($x_i!$ ~ const wrt λ)

$$= \arg\max_{\lambda} e^{-n\lambda}\lambda^{\Sigma x_i}$$

Since log is increasing fn $\arg\max_{\lambda} p_\lambda(x_1 \cdots x_n) = \arg\max_{\lambda} \log p_\lambda(x_1 \cdots x_n)$

$$= \arg\max_{\lambda} -n\lambda + \Sigma x_i \log \lambda$$

To maximize set $\frac{\partial}{\partial \lambda} = 0$ giving

$$0 = -n + \frac{\Sigma x_i}{\lambda} \Rightarrow \hat{\lambda} = \frac{\Sigma x_i}{n} \doteq \bar{x}$$

the "sample mean"

Comment

When dist parametrized by $\theta =$ mean (as with Poisson) often happens that $\hat{\theta}_{MLE} = \bar{x}$

Ex (Normal with known σ^2)

Have $X_1 \cdots X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ where σ^2 known, μ unknown.

$$\hat{\mu}_{MLE}(x_1 \cdots x_n) = \arg\max_{\mu} q_\mu(x_1 \cdots x_n) = \arg\max_{\mu} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i - \mu)^2}$$

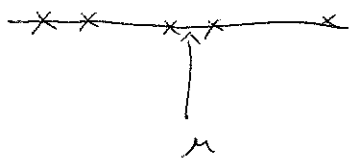
$$= \arg\max_{\mu} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2}$$

$$= \arg\max_{\mu} \log e^{-\frac{1}{2\sigma^2}\sum(x_i - \mu)^2}$$

$$= \arg\max_{\mu} -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$= \arg\min_{\mu} \sum_{i=1}^{n}(x_i - \mu)^2$$

$$= \bar{x}.$$



Q: What is # that minimizes sum sq diff from set of #'s $x_1 \cdots x_n$?

A: $\bar{x}$

$$\sum_{i=1}^{n}(x_i - \mu)^2 = \sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - \mu)^2$$

$$= \underbrace{\sum_{i=1}^{n}(x_i - \bar{x})^2}_{\text{const wrt } \mu} + \sum(\bar{x} - \mu)^2 + 2\sum(x_i - \bar{x})(\bar{x} - \mu)$$

$(x_i - \bar{x})$ sums to 0

To minimize $\sum_{i=1}^{n}(x_i - \mu)^2$ take $\mu = \bar{x}$.

Ex (Normal with known μ unknown σ^2)

$$\hat{\sigma}^2_{MLE}(x_1 \cdots x_n) = \arg\max_{\sigma^2} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}$$

$$= \arg\max_{\sigma^2} (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

$$= \arg\max_{\sigma^2} -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum(x_i-\mu)^2$$

Setting $\frac{d}{d\sigma^2} = 0$ gives

$$0 = \frac{-n/2}{\sigma^2} + \frac{1}{2}\sum(x_i-\mu)^2 \frac{1}{(\sigma^2)^2}$$

$$\Rightarrow \quad \hat{\sigma}^2_{MLE} = \frac{1}{n}\sum(x_i-\mu)^2 \qquad (= \text{the "sample" ave sq dist. from } \mu)$$

Bayesian Estimation

As before: have $p_\theta(x)$ for unknown θ. Observe x $(= x_1 \cdots x_n)$ and want to estimate θ.

Bayesian View Assume prior dist on θ, $p(\theta)$. Then

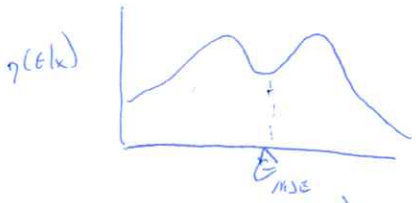
$$p(\theta|x) = \frac{p(\theta)\, p(x|\theta)}{p(x)}$$

Will base estimate for θ on $p(\theta|x)$. Eg

① $\hat{\theta}_{MAP} = \arg\max_\theta p(\theta|x) = \arg\max_\theta p(\theta)\, p(x|\theta)$

Maximum A Posteriori

② $\hat{\theta}_{MSE} = E(\theta|x)$



(Recall that mean = expectation minimizes our sq. error)

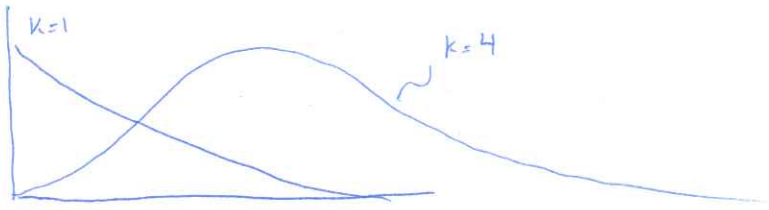
$$\left[\arg\min_c \int (x-c)^2 p(x)\,dx = \int x\, p(x)\,dx = \mathbb{E}X \right]$$

Conjugate Priors

Gamma Dist

$x \geq 0$

$$X \sim \text{Gamma}(\theta, k) \qquad p(x) = \frac{e^{-x/\theta} x^{k-1}}{\Gamma(k)\, \theta^k} \qquad \Gamma(k) = \int_0^\infty e^{-x} x^{k-1}\,dx$$



Interesting interaction between Gamma and Poisson.

Observe $x \sim \text{Poisson}(\lambda)$ and want to estimate λ.

Assume prior on λ, $\lambda \sim \text{Gamma}(\theta, k)$

$$p(\lambda|x) = \frac{p(\lambda)\, p(x|\lambda)}{p(x)} = \frac{e^{-\lambda/\theta}\lambda^{k-1}}{\Gamma(k)\theta^k} \frac{e^{-\lambda}\lambda^x}{x!} \frac{1}{p(x)}$$

denom is const. wrt λ

$$= c\, e^{-\lambda / \frac{1}{1+1/\theta}} \lambda^{x+k-1}$$

$$\Rightarrow \quad \lambda|x \sim \text{Gamma}\left(\frac{1}{1+1/\theta},\ x+k\right)$$

The parameter update after observation is

$$\theta_{new} = \frac{1}{1 + 1/\theta_{old}}$$

$$k_{new} = x + k_{old}.$$

Note: We ended up where we started with a Gamma Dist on λ

Could observe another $x, x_2 \sim \text{Poisson}(\lambda)$, then

$$\lambda | x_1, x_2 \sim \text{Gamma}\left(\frac{1}{2+1/\theta},\ x_1+x_2+k\right)$$

$$\lambda | x_1, x_2, x_3 \sim \text{Gamma}\left(\frac{1}{3+1/\theta},\ x_1+x_2+x_3+k\right)$$

$$\lambda | x_1 \cdots x_n \sim \text{Gamma}\left(\frac{1}{n+1/\theta},\ \sum_{i=1}^{n} x_i + k\right)$$

Do earthquakes.r

Conjugate Priors Cont

Have $X \sim N(\mu, \sigma^2)$ and want to estimate μ.

Have prior dist on μ, $\mu \sim N(r, \rho^2)$

What is posterior on μ?

$$g(\mu|x) = \frac{g(\mu)\, g(x|\mu)}{g(x)} = c\, e^{\frac{-1}{2\rho^2}(\mu - r)^2} e^{\frac{-1}{2\sigma^2}(x-\mu)^2}$$

$$= c' e^{q(\mu)}$$

where $q(\mu) = a\mu^2 + b\mu + c = \frac{1}{\beta^2}(\mu - \alpha)^2 + \gamma$

By completing the square get

$$\mu | x \sim N\left(\frac{\frac{r}{\rho^2} + \frac{x}{\sigma^2}}{\frac{1}{\rho^2} + \frac{1}{\sigma^2}}, \frac{1}{\frac{1}{\rho^2} + \frac{1}{\sigma^2}} \right)$$

Said another way, let

r_{old} ρ^2_{old} be original (prior) parms for μ

r_{new} ρ^2_{new} " new (posterior) " " "

$$r_{new} = \frac{\frac{r_{old}}{\rho^2_{old}} + \frac{x}{\sigma^2}}{\frac{1}{\rho^2_{old}} + \frac{1}{\sigma^2}}$$

$$\rho^2_{new} = \frac{1}{\frac{1}{\rho^2_{old}} + \frac{1}{\sigma^2}}$$

As before can use this update eqn. to revise our knowledge of μ with sequence $X_1 \cdots X_n \overset{iid}{\sim} N(\mu, \sigma^2)$.

Continued in HW.

Generative vs. Discriminative Classification

Bishop Ch. 4

In generative model have $p(c)$, $p(x|c)$ and compute

$$p(c|x) = \frac{p(c)\,p(x|c)}{\sum_{c'} p(c')\,p(x|c')}$$

posterior is everything we need to know for inferences about c.

Typically involves

① Estimating $p(c)$ ~ Eg $N(\mu_c, \Sigma_c)$

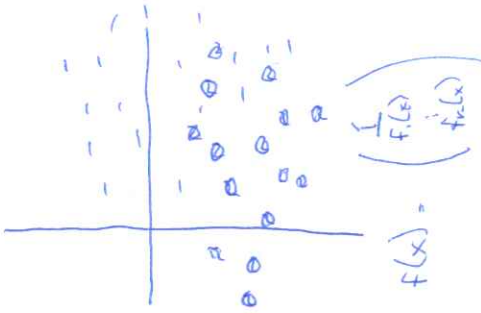
② Estimating parameters for $p(x|c)$

Discriminative classification models $p(c|x)$ directly (and parametrically)

Most famous example: Logistic regression.

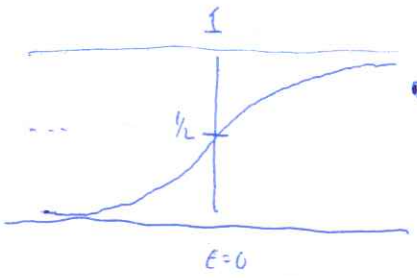
Logistic Regression

Consider 2-class classification problem



Define logistic sigmoid fn

$$\sigma(\epsilon) = \frac{1}{1+e^{-\epsilon}}$$



Important properties

① $\sigma(\epsilon) + \sigma(-\epsilon) = 1$

② $\sigma'(\epsilon) = \sigma(\epsilon)\,(1-\sigma(\epsilon))$

Model

Have observation $x = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}$

Compute features: $f_1(x) \dots f_K(x)$

Model

$$p(c=1|x) = \sigma\left(w_0 1 + w_1 f_1(x) + \cdots + w_K f_K(x)\right)$$

$f(x)$ = "feature vector" contains whatever observables, or functions of observable we think are useful,

Ex

2 classes $c_i = 1$ has disease

$c_i = 0$ no disease

Measure 3 important biomarkers x_1, x_2, x_3.

Believe interactions between vbles also important so let

$$f(x) = (x_1, x_2, x_3, x_1x_2, x_2x_3, x_1x_3, x_1x_2x_3, 1)$$

$$p(c=1 \mid x) = \sigma(w^t f(x)) = \sigma(w_0 + w_1x_1 + w_2x_2 + w_3x_3 + w_4x_1x_2 + \dots)$$

hw 2 sols

Notes

Practice problems

Change due date

Logistic assumes $p(c=1 \mid x) = \sigma(w^t f(x))$

but how to find $w = (w_0 \cdots w_k)^t$?

① Since $p(c=1 \mid x) = \sigma(w^t f(x))$

$p(c=0 \mid x) = 1 - \sigma(w^t f(x))$

② If $c \in \{0,1\}$ could write as

$$p(c \mid x) = \sigma(w^t f(x))^c \left(1 - \sigma(w^t f(x))\right)^{1-c}$$

③ Suppose have sample $(c_1, x_1), (c_2, x_2) \cdots (c_n, x_n)$

Assume the ~~samples~~ observations independent.

$$p(c_1 \cdots c_n \mid x_1 \cdots x_n) = \prod_{i=1}^{n} p(c_i \mid x_i) = \prod_{i=1}^{n} \sigma(w^t f(x_i))^{c_i} (1 - w^t f(\cdots)^{1-c_i}$$

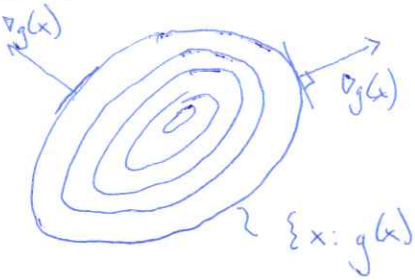
Choose w by max like. That is

$$\hat{w} = \arg\max_w \prod_{i=1}^{n} p(c_i \mid x_i) = \arg\max_w \prod_{i=1}^{n} \sigma(w^t f(x_i)) \left(1 - \sigma(w^t f(x_i))\right)^{1-c_i}$$

Can't optimize in closed form as in previous examples, but can optimize numerically.

Recall

Want to maximize / minimize $g(x) : \mathbb{R}^n \to \mathbb{R}$



$\nabla g(x) = \begin{pmatrix} \frac{\partial g(x)}{\partial x_1} \\ \vdots \\ \frac{\partial g(x)}{\partial x_n} \end{pmatrix}$ is gradient.

4) $\nabla g(x) = 0$ at optimum or saddle

1) $\nabla g(x)$ "points" in direction of maximal increase

2) $-\nabla g(x)$ " " " decrease

3) $\nabla g(x)$ orthogonal to level sets

Back to Logistic Regression

$$q(c_1 \cdots c_n | x_1 \cdots x_n) = \prod_{i=1}^{n} \sigma(w^t f(x_i))^{c_i} \left(1 - \sigma(w^t f(x_i))\right)^{1-c_i}$$

$$\log q(c_1 \cdots c_n | x_1 \cdots x_n) = \sum_{i=1}^{n} c_i \log \sigma(w^t f(x_i)) + (1-c_i) \log\left(1 - \sigma(w^t f(x_i))\right)$$

$$\frac{d \log q(c_1 \cdots c_n | x_1 \cdots x_n)}{d w_k} = \sum_{i=1}^{n} \frac{c_i \sigma'(w^t f(x_i)) f_k(x_i)}{\sigma(w^t f(x_i))} - \frac{(1-c_i)\sigma'(w^t f(x_i)) f_k(x_i)}{1 - \sigma(w^t f(x_i))}$$

$\sigma'(t) = \sigma(t)(1-\sigma(t))$

$$= \sum c_i \left(1 - \sigma(w^t f(x_i))\right) f_k(x_i) - (1-c_i)\sigma(w^t f(x_i)) f_k(x_i)$$

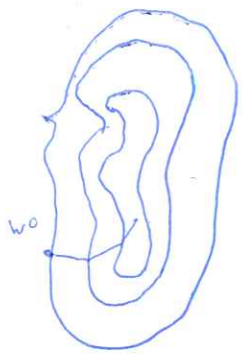
$$= \sum_{i=1}^{n} \left(c_i - \sigma(w^t f(x_i))\right) f_k(x_i)$$

$$\Rightarrow \nabla \log q(c_1 \cdots c_n | x_1 \cdots x_n) = \sum_{i=1}^{n} \left(c_i - \sigma(w^t f(x_i))\right) \begin{pmatrix} \vdots \\ f(x_i) \\ \vdots \end{pmatrix}$$

Can optimize $\log q(c_1 \cdots c_n | x_1 \cdots x_n)$ by gradient descent

Choose initial guess for w, w^0

$$w^{new} = w^{old} + \varepsilon \nabla \log q(c_1 \cdots c_n | x_1 \cdots x_n) \quad \text{for some stepsize } \varepsilon$$

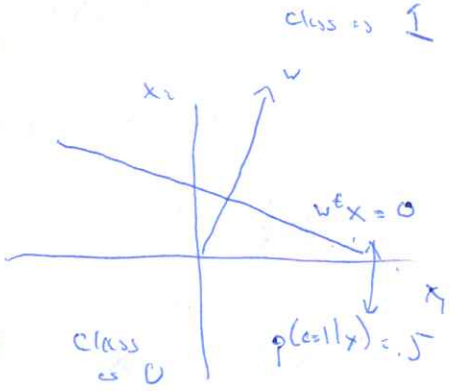


Continue until ∇ sufficiently small

(at optimal value $\nabla = 0$)

① Do logistic_model.r

② logistic_grad_descent.r



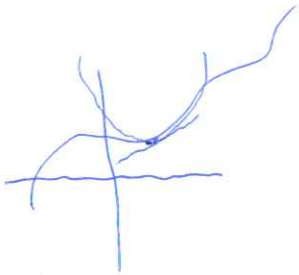
Newton - Raphson Method

NR improves on simple-minded gradient descent. $f: \mathbb{R}^n \to \mathbb{R}$

$$x^{new} = x^{old} + \varepsilon \nabla f(x^{old})$$

Recall Taylor series in 1-d $f: \mathbb{R} \to \mathbb{R}$

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2$$



The multivariate T.S. is $f: \mathbb{R}^n \to \mathbb{R}$

$$f(x) \approx f(x_0) + \underbrace{\nabla f(x_0) \cdot (x - x_0)}_{\text{dot prod.}} + \frac{1}{2} (x - x_0)^t H (x - x_0)$$

where H is "Hessian" matrix $\underset{n \times n}{H_{ij}} = \dfrac{\partial^2 f(x_0)}{\partial x_i \partial x_j}$

Interesting fact on quadratic forms

$$q(x) = \frac{1}{2} (x - x_0)^t H (x - x_0) \qquad \nabla q(x) = H(x - x_0)$$

Take gradient of approx above to get

$$\nabla f(x) \approx \nabla f(x_0) + H(x - x_0)$$

To find ~~optimal~~ maximal value of $f(x)$ set $\nabla f(x) = 0$ to get

$$0 = \nabla f(x) \approx \nabla f(x_0) + H(x - x_0)$$

Solving for x gives

$$-H \nabla f(x_0) = x - x_0 \implies x^{opt} = x_0 - H \nabla f(x_0)$$

NR: $$x^{new} = x^{old} - H \nabla f(x^{old})$$

Logistic Regression with Newton-Raphson

Review

$$\sigma(t) = \frac{1}{1+e^{-t}} \qquad \text{Model} \quad p(c=1 \mid x) = \sigma(w^t x)$$

Have data vectors $x_1, x_2, \dots x_n$ We know

$$\frac{\partial \log p(c_1 \dots c_n \mid x_1 \dots x_n)}{\partial w_j} = \sum_{i=1}^{n} (c_i - \sigma(w^t x_i)) x_{ij} \iff \nabla \log p(c_1 \dots c_n \mid x_1 \dots x_n) = \sum (c_i - \sigma(w^t x_i)) x_i$$

Writing $X_{n \times k} = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ x_{21} & \cdots & x_{2k} \\ \vdots & & \\ x_{n1} & \cdots & x_{nk} \end{pmatrix}$ $\quad c = \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix}$ $\quad p = \begin{pmatrix} \sigma(w^t x_1) \\ \vdots \\ \sigma(w^t x_n) \end{pmatrix}$

$$\nabla \log p(c_1 \dots c_n \mid x_1 \dots x_n) = X^t (c - p)$$

$$\frac{\partial^2 \log p(c_1 \dots c_n \mid x_1 \dots x_n)}{\partial w_j \partial w_l} = -\sum_{i=1}^{n} \sigma(w^t x_i)(1 - \sigma(w^t x_i)) x_j x_l$$

$$= -(X^t C X)_{jl}$$

$$C = \begin{pmatrix} \sigma(w^t x_1)(1-\sigma(w^t x_1)) & & 0 \\ & \ddots & \\ 0 & & \sigma(w^t x_n)(1-\sigma(w^t x_n)) \end{pmatrix}$$

ie. $-X^t C X$ is Hessian

$$NR: \quad w^{new} = w^{old} + H^{-1} \nabla f(w^{old}) = w^{old} + (X^t C X)^{-1} X^t (c - p)$$

[NB. C and p depend on w^{old} so must be recomputed each iter