1 HG48600/STAT34550 Lectures

1.1 Lecture 1: Introduction to Course and Probability

Introduction

- Themes of this course:
  - Thinking probabilistically
    - The systems we aim to model are inherently **stochastic**
    - Probabilities gives us a language for expressing our uncertainy in precise terms (i.e. we are often going to be thinking as Bayesians)
  - Handling complex probability distributions
    - Those with an index set (i.e. **stochastic processes**)
    - **Heirarchical models** with underlying **latent (hidden)** variables
  - Constructing custom solutions to inference problems in biology
    - Recognizing the biological aspects of a problem and being able to build it into our solutions, i.e. not being beholden to fitting a problem into frameworks already invented
    - That said, we will learn several general purpose models
- Broader context for this course
  - We see three domains are commonly mastered by the best computational biologists.
  - This course will cover 2 of them at an introductory level: Stochastic processes and inference in complex, heirarchical models.
  - The third domain will be the subject of a course that will be taught next year: Computational data structures and algorithms.

Course expectations

- Problem Sets
  - 5 total: You will have at least 1 week to complete them
- Final project
  - Do something interesting leveraging the concepts of this course
  - Use ideas from this course to address a small problem in an area of biology that interests you (need not be your PhD research area)
  - Develop a teaching vignette / lab for a subject area of this course
  - Poster Session on the last day of class
- Scribe duty:
  - You will take notes, most likely on pen and paper.
  - After class you will write them up via latex (or markdown) and post.
  - Please sign up with Evan.

Review: Marginal, Joint, and Conditional distributions, Bayes Rule

- Motivation
  - Most problems we work on involve multiple random variables.
  - To think about multiple random variables at a time it is useful to understand **joint**, **marginal** and **conditional** distributions. There are also analogous forms for expectations, variances, and covariances.
- Example: A basic two-variable discrete joint probability distribution
  - Example 1

X|Y	$Y = 1$	$Y = 2$	$P(X = x)$
$X = 0$	0.08	0.12	0.2
$X = 1$	0.16	0.24	0.4
$X = 2$	0.12	0.18	0.3
$X = 3$	0.04	0.06	0.1
$P(Y = y)$	0.4	0.6	

- Conditional probability and independence:
  - The basic definition
$$P(B|A) = \frac{P(A,B)}{P(A)}$$
    Note: Trivially generalizes for talking about discrete or continuous random variables.
    Also note: we like to replace the formal notation $P(A = a)$ by $P(A)$.
  - Independence
    * Two events A,B are said to be indepdent if $P(A,B) = P(A)P(B)$
    * Note from def of conditional probablity this implies: $P(B|A) = P(B)$ (and $P(A|B) = P(A)$)
    * A big theme of the course will be leveraging conditional probabilities and independence to solve problems.
  - Marignal distributions and the law of total probability: We can "marginalize" by a summation operation:
$$P(A = a) = \sum_{b:P(B=b)>0} P(A = a, B = b)$$
    or
$$P(A = a) = \sum_{b:P(B=b)>0} P(A = a|B = b)P(B = b)$$
    or in shorthand
$$P(A) = \sum P(A|B)P(B)$$
    Note: As is often the case, the analogous form for continous random variables replaces the summation step with integration.
- Bayes' rule
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$
  This has tremendous utility as a tool for taking one conditional probablity ($P(B|A)$) and computing it's "inverse" $P(A|B)$. It also has great utility for inference problems and shows up in the following form. (Matthew will expand on this latter point)
$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} = fracP(X|\theta)P(\theta)\int P(X|theta)P(\theta)d\theta$$
  Where, X are some data, and θ are the parameters of our model.

Review: Introduction to Random Variables

- Basic definitions:
  - Ω : The sample space; points in Ω represent **elementary events**
  - Probability:
    - A function that ascribes a measure to each point (and subset of points) in the sample space, with the important property that the integral of the measure over Ω equals 1.
    - Interpretations: The frequency at which an event will occur, a measure of uncertainty
  - Random variables : Real-valued function over the elementary events in the sample space.
    - Example: X is the sum of two fair die.
      - $X = 2$ if the first die is 1 and the second is 1.
    - Example: An **indicator variable** for whether a single die is even.
      - $I_{odd} = 1$ if die role is single die role is 2, 4, 6; and 0 otherwise.
    - Probabilities can be assigned to the values of random variables
    - Typically we think at the level of random variables and probability distributions/densities (and ignore the more formal construction of the sample space and measure definitions)

- Basic Discrete Random Variables:

Name	parameters	probability mass function	Mean	Variance
Binomial	$n > 0$ and $0 \leq p \leq 1$	$\binom{n}{x} p^x (1-p)^x$	np	$np(1-p)$
Poisson	$\lambda > 0$	$e^{-\lambda} \frac{\lambda^x}{x!}$	λ	λ
Geometric	$0 \leq p \leq 1$	$p(1-p)^{x-1}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$

  See Ross Table 2.1

- Basic Continuous Random Variables:

Name	parameters	probability density function	Mean	Variance
Uniform	a,b	$\frac{1}{b-a}$ for $a < x < b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential	$\lambda > 0$	$\lambda e^{-\lambda x}$ for $x > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma	n>0,$\lambda > 0$	$\frac{\lambda e^{-\lambda x}(\lambda x)^{n-1}}{(n-1)!}$ for $x \geq 0$	$\frac{n}{\lambda}$	$\frac{n}{\lambda^2}$
Normal	μ, $\sigma^2 > 0$	$\frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
Beta	$\alpha > 0$,$\beta > 0$	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

  - Note: See Ross Table 2.2

- Additional random variable distribution that will be of interest for this course
  - Distributions of the exponential family, in particular:
    - Multinomial distribution
    - Dirichlet distribution (a multivariate analog of the beta)
    - Multivariate Normal distribution

- Definition of a stochastic process
  - We will spend a large amount of our time thinking about a special collection of random variables known as a **stochastic process**
  - A stochastic process is a set: $X(t), t \in T$
  - $X(t)$ as the **state** of the system at time t.
  - T as the **index set** of the process. t often interpreted as time variable or a spatial variable.

- **State space** : The set of possible values of $X(t)$
- Stochastic processes are a family of random variables that describe the evolution through time of some (physical) process.
- We will use stochastic processes as models for biological processes, and as a trick to simulate from intractable distributions (this is the idea of MCMC and Gibbs sampling).

Review: Expectation, Variances, Covariances

- Definition of Expectation
  - Discrete case:
$$E[X] = \sum_{x:p(x)>0} xp(x)$$
  - Continous case:
$$E[X] = \int_{-\infty}^{\infty} xf(x)dx$$
  - Expectations of functions
    - g(X) is itself a random variable.
    - In simple cases, $E[g(X)]$ can be computed from $E[X]$. For example:
      - $E[aX + b] = aE[X] + b$
    - In more complicated cases we would have to compute the integral $\int g(x)f(x)dx$, or the discrete analog.
  - Another way to calculate expectations:
$$E[X] = \int_0^{\infty} [-F(-x) + (1 - F(x))]$$
- Definition of variance
$$Var(X) = E[(X - E[X])^2] = E[X^2] - E[X]^2$$
- Definition of covariance
  - Definition
$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$
  - If X,Y are independent, covariance equals 0.
  - Useful result:
$$Var(aX + bY + c) = a^2Var(X) + b^2Var(Y) + 2abCov(X, Y)$$

The Law of Large Numbers and introduction to Monte Carlo

- **The Strong Law of Large Numbers**: Let $X_1, X_2, \ldots$ be a sequence of independent, identically distributed variables, and let $E[X_i] = \mu$ (where μ is finite). Then,
$$P(lim_{n\to\infty} \frac{X_1 + X_2 + \ldots + X_n}{n} = \mu) = 1$$
- This result forms the basis of "vanilla" Monte Carlo estimators:
  - For expectations:
$$E[g(X)] \approx \frac{1}{M} \sum_{i=1}^{M} g(x_i)$$
    where $x_i \sim f_X(\cdot)$

– For probabilities (using indicator functions):

$$P(X = x) = E[I_{X=x}] \approx \frac{1}{M}\sum_{i=1}^{M} I_{X=x}(x_i)$$

where $x_i \sim f_X(\cdot)$

– Thus by being able to simulate instances of a random variable X we can compute probabilities of events dependent on X as well as computing expectations that require integrating over all possible values of X.

– This "Monte Carlo" strategy is a workhorse of modern computational statistics. It also has many variants, several of which we'll learn about in the course (e.g. Gibbs, MCMC).

Conditional expectations and variances

- Definition of Conditional Expectation
  - Discrete case:

$$E[X|Y = y] = \sum_x xP(X = x|Y = y) = \sum_x xp_{X|Y}(x|y)$$

    where $p_{X|Y}(x|y) = p(x, y)/p_Y(y)$
  - Continuous case:

$$E[X|Y = y] = \int_{\infty}^{\infty} xf_{X|Y}(x|y)dx$$

    where $f_{X|Y}(x|y) = f(x, y)/f_Y(y)$.
  - Note:
    * Simple, it's just an expectation over a conditional distribution/density function.
    * And note, $E[X|Y = y]$ is a random variable that is a function of y. Thus we can compute it's expectation: $E[E[X|Y]]$. This turns out to be very useful…
- Computing Expectations, Variances and Probabilities by Conditioning
  - Computing expectations of conditional expectations gives us a new route to computing an expectation (**Law of total expectation**):

$$E[X] = E[E[X|Y]]$$

  - We can also compute variances (**Law of total variance**):

$$Var(X) = E[Var(X|Y)] + Var(E[X|Y])$$

  - And for computing probabilities (using indicator variables)

$$I_E = \begin{cases} 1 & \text{E happens} \\ 0 & \text{otherwise} \end{cases}$$

$$E[I_E] = 1P(I_E = 1) + 0P(I_E = 0) = P(E)$$

- Examples of using conditioning to compute probabilites:
  - Ross Example 3.10 and 3.19 : Mean and Variance of a Compound Variable

- Example 3.10: Expected number of accidents in a week is 4 and the number of workers injured in each accident is an indpt RV with mean 2. What is the number of expected injuries during a week?

  Solution: Let N denote the number of accidents, and X_i the number injuries per accident. Our interest is:
  $$E[\sum_1^N X_i] = E[E[\sum_1^N X_i|N]]$$
  Note:
  $$E[\sum_i^n X_i|N = n] = E[\sum_1^n X_i] = nE[X]$$
  and then plugging in get:
  $$E[E[\sum_1^n X_i|N]] = E[nE[X]] = E[N]E[X]$$
  This is kind of obvious but now we've been rigorous about it. More interestingly, what about the variance?
- Example 3.19: Let S be the compound variable $\sum_1^N X_i$. Find the variance. Let $Var(X) = \sigma^2$ and $E[X] = \mu$. We'll use the conditional variance formula.

  Solution:
  $$Var(S) = E[Var(S|N)] + Var(E[S|N])$$
  First term:
  $$Var(S|N = n) = Var(\sum_{i=1}^n X_i) = n\sigma^2$$
  $$E[Var(S|N)] = E[N]\sigma^2$$
  Second term:
  $$E[S|N] = n\mu$$
  $Var(E[S|N])$ then equals $\mu^2 Var(N)$

  So we have: Var(S)=σ^2 E[N]+μ^2 Var(N). In special case where N is Poisson(λ) we have:
  $$Var(S) = \lambda\sigma^2 + \lambda\mu^2$$
  which note has the simplification: $\lambda E[X^2]$.

Conclusions for the day

- For working on probablity problems...
  - Conditioning often helps
  - Use indicator variables to your advantage
  - Train yourself to recognize probablity distributions when they appear (as in Example 3.23 with the Poissons appear)
  - Sometimes its useful to remember distributions sum (or integrate to 1) (see Ross 3.22 for an example with a Gamma that appears in the simplified form).
  - Use tools from "real analysis":
    - Recognize that many ugly looking sum's or integrals have analytic solutions (e.g. see Example 3.25 or section 3.63). Mathematica can help recognize these
    - Proofs using induction are often needed. Similarly, recursive formulas often arise and can be solved (Example 3.26).

- Advanced:
  - Using probablistic inequalities to form bounds
  - Using moment generating functions and characteristic functions for solving problems with sums of random variables

Miscellaneous Review

- Cumulative distribution functions and density functions
  - Cumulative distribution function: $F(b) = P(X <= b)$
    - F(b) is non-decreasing in b
    - $\lim_{b \to \infty} = F(\infty) = 1$
    - $\lim_{b \to \infty} = F(-\infty) = 0$
    - CDF's take the form of step functions for discrete RVs
    - For continuous RV's
      - $F(a) = P(X \in (\infty, a)) = \int_{-\infty}^{a} f(x)dx$
      - $\frac{d}{da}F(a) = f(a)$, ie density is the derivative of the cdf
- Definition of Covariance

$$E[X,Y] = E[XY] - E[X]E[Y]$$

  Properties of covariance:
  - $Cov(X,X) = Var(X)$
  - $Cov(X,Y) = Cov(Y,X)$
  - $Cov(cX,Y) = cCov(X,Y)$
  - $Cov(X,Y+Z) = Cov(X,Y) + Cov(X,Z)$
- The **Chain Rule**
  - In its basic form:

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

  - Which generalizes as:

$$P(A_1, A_2, \ldots, A_k) = P(A_1)P(A_2|A_1)\ldots P(A_k|A_{k-1})$$

  - This result holds regardless of the ordering.