STA561: Probabilistic machine learning

Concepts in Probability and Statistics (8/26/13)

Lecturer: Barbara Engelhardt *Scribes: Ben Burchfiel, Nicole Dalzell, Andrew Kephart, Florian Wagner*

Course Outline

- **Course Website:** http://www.genome.duke.edu/labs/engelhardt/courses/sta561.html
  - **helpful resources:** tutorials, videos, additional readings
  - syllabus
  - office hours
- **Book:** Kevin Murphy's 'Machine Learning: a probabilistic perspective'
  - readings are highly recommended, due to the fast pace of this course
- **Homework:** approx. weekly; due in one week
  - study groups allowed, but write solutions independently
  - include names of all group members on homework
  - submit in a single pdf to Sakai website before class.
  - regrades done by Professor, does not mean higher grade
- **Scribe Notes:** done in LaTeX
  - sign up sheet online
  - four scribes per class
- **Take home Midterm:** Due Date: October 16
- **Final Project**
  - 1 page proposal due October 2nd.
  - 4 page paper due on December 4th.
  - Poster Presentation: Either December 4th or December 14th, 2-5pm
  - Work alone or in pairs, but pairs held to higher standard
  - Final exam (if and only if you are taking this class for CS AI/ML quails credit): Either December 4th or December 14th, 2-5pm

Why Machine Learning

Fundamental Challenge: It is easy to collect data, but it is hard to extract meaning from it.

- Machine Learning (ML) gives you the tools necessary to extract meaning from massive data.
- The industry is expanding rapidly. There are more data than analytic solutions.
- Duke is one of the best places to study ML because we have:
  - one of the best Bayesian Statistics departments in the world
  - a great ML community with
    - seminars Wednesdays at 3:30
    - lunch seminars for students

Probability

Often the goal of machine learning (or research in general) is to determine the probability of an event, e.g.:

- $\Pr(\text{year that polar ice cap melts} \leq 2020)$
- $\Pr(\text{a new email is spam})$
- $\Pr(\text{a person is at risk for a disease})$

There are two main paradigms in statistics:

- **Frequentist:** probabilities are long run frequencies
  - flip a coin a million times to determine if it's fair
- **Bayesian:** probabilities quantify our uncertainty in events
  - designed to get the closest to the truth given a specific set of data

From the Frequentist perspective, the probability of an event is defined as the *frequency* that specific event occurred in a long list of repeated trials.

In ML, we are working with the data that we are given, which often means that we do not have the luxury of observing multiple trials. We therefore require the ability to generalize from a small number of events. This is easily described through the Bayesian paradigm. We will adopt the Bayesian, with a number of exceptions, in our ML course.

1 Probability Theory

A *random variable* (RV) describes an event that we have not yet observed, e.g.:

- the number of heads in 6 flips of a fair coin,

- the number of spam e-mails we will receive today,
- whether it rains today.

Let $\mathsf{Pr}(A) \triangleq$ the probability that an event, A, occurs. Then, $0 \leq \mathsf{Pr}(A) \leq 1$ and $\mathsf{Pr}(\overline{A}) = 1 - \mathsf{Pr}(A)$.

$\overline{A}$ represents "not A," or the event A not happening. More formally, $\overline{A}$ is called the *complement* of A, and $\mathsf{Pr}(\overline{A})$ can be interpreted as the probability that event A does not occur. When we say that X is a random variable, and we write $\mathsf{Pr}(X = x)$ this is the probability that RV X takes on value x.

A *binary random variable* is a RV that can take on only two values, e.g., 0 or 1. Suppose A is a binary RV. We may then write $A \in \{0, 1\}$, meaning that A can take on the values 0 or 1.

A simple example of a situation involving a a binary random variable is the flipping of a fair coin one time. Will the outcome of the flip be heads (1) or tails (0)? If we let A denote the unknown outcome of the flip, then A is a binary random variable.

Discrete Random Variables

Discrete random variables have a discrete sample space, meaning that they can take on only a finite number of values. Suppose $\chi = \{1, 2, 3, 4\}$ is our sample space, and X is a random variable that takes values in this sample space, ie $X \in \chi$. Consider, for instance, reaching into a bag with 4 balls, with labels 1, 2, 3 and 4. What is the label on the ball that you select?

As before $0 \leq \mathsf{Pr}(X) \leq 1$; If $\mathsf{Pr}(X = i) = 0$, you never draw ball i. Alternatively, if $\mathsf{Pr}(X = i) = 1$, you always draw ball i. In all other cases, $(0 < \mathsf{Pr}(X = i) < 1)$, you sometimes draw ball i. Additionally, the ball that you draw must have either a 1,2,3 or 4 on it, so

$$\mathsf{Pr}(X = 1) + \mathsf{Pr}(X = 2) + \mathsf{Pr}(X = 3) + \mathsf{Pr}(X = 4) = 1$$

More generally, $\sum_{x \in \chi} \mathsf{Pr}(X = x) = 1$.

1.1 Rules of Probability

Consider two events A and B.

Union

$$\mathsf{Pr}(A \text{ or } B) \triangleq \mathsf{Pr}(A \cup B) = \mathsf{Pr}(A) + \mathsf{Pr}(B) - \mathsf{Pr}(A \cap B).$$

Note that when we add $\mathsf{Pr}(A)$ and $\mathsf{Pr}(B)$, we double count the probability of the intersection of these events $\mathsf{Pr}(A \cap B)$.

Joint Probability

Intersection

$$\mathsf{Pr}(A \text{ and } B) \triangleq \mathsf{Pr}(A \cap B) \triangleq \mathsf{Pr}(A, B) = \mathsf{Pr}(A \mid B)\mathsf{Pr}(B).$$

Conditional Probabilities

$\mathsf{Pr}(A \mid B)$ means "the *conditional* probability of A given B", e.g. "What is the probability that it will rain, *given* that it is 79 degrees today?" We can define the conditional probability in terms of the joint probability and the marginal probability:

$$\mathsf{Pr}(A \mid B) = \frac{\mathsf{Pr}(A, B)}{\mathsf{Pr}(B)}$$

Chain Rule

If C and D also denote events,

$$\mathsf{Pr}(A \text{ and } B \text{ and } C \text{ and } D) \triangleq \mathsf{Pr}(A \cap B \cap C \cap D) \triangleq \mathsf{Pr}(A, B, C, D) = \mathsf{Pr}(A)\mathsf{Pr}(B \mid A)\mathsf{Pr}(C \mid A, B)\mathsf{Pr}(D \mid A, B, C)$$

Marginal Probability

If we are interested in the *marginal probability* $\mathsf{Pr}(A)$, and we have information on the joint $\mathsf{Pr}(A, B)$, we may obtain information on the marginal probability by *marginalizing* over B. In the case of discrete random variables, we do this by summing over B.

$$\begin{aligned} P(A) &= \sum_{b \in \mathcal{B}} \mathsf{Pr}(A, B = b) \\ &= \sum_{b \in \mathcal{B}} \mathsf{Pr}(A \mid B = b)\mathsf{Pr}(B = b) \end{aligned}$$

Bayes Rule

Let's return to the idea of conditional probability for a moment. From the definition of joint probability, we can solve for the conditional probability of $A \mid B$ as long as $\mathsf{Pr}(B) > 0$.

$$\begin{aligned} \mathsf{Pr}(A, B) = \mathsf{Pr}(A \mid B)\mathsf{Pr}(B) \Rightarrow \mathsf{Pr}(A \mid B) &= \frac{\mathsf{Pr}(A, B)}{\mathsf{Pr}(B)} \\ \Rightarrow \mathsf{Pr}(A \mid B) &= \frac{\mathsf{Pr}(B \mid A)\mathsf{Pr}(A)}{\mathsf{Pr}(B)} \end{aligned}$$

The final (blue) equation is called *Bayes rule* and will be used extensively in the derivation of Bayesian posterior distributions.

Continuous Random Variables

In the discrete case, a random variable X could take on only finitely many values. In the continuous case, the random variable may take on infinitely many values. Let's begin with an example, a Uniform Random Variable.

For continuous random variables the function $\mathsf{p}(\cdot)$ is called a *probability density function* (PDF).

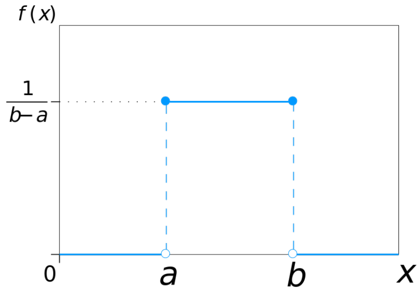

Figure 1: Uniform(a,b). Source: Wikipedia

We say that the Uniform random variable X has support on the set (a, b), meaning that $a < X < b$ (Figure **??**). Over that support, the pdf of x is defined as $\mathsf{p}(x) = \frac{1}{b-a} > 0$, and outside of the support, $\mathsf{p}(x) = 0$. Note that the pdf $\mathsf{p}(x)$ can be bigger than 1, but it always integrates to exactly one. $\{a = 0, b = 0.2 \Rightarrow \mathsf{p}(0.1) = \frac{1}{0.2} = 5\}$

Let's consider a few events to illustrate this point:

Event A	$a \leq X < \frac{b+a}{2}$	$\mathsf{Pr}(A) = 1/2$
Event B	$\frac{b+a}{2} \leq X < b$	$\mathsf{Pr}(B) = 1/2$
Event W	$a \leq X \leq b$	$\mathsf{Pr}(W) = 1$

Events A and B are *mutually exclusive*, meaning that they cannot happen at the same time. X cannot be both $< \frac{b+a}{2}$ and $\geq \frac{b+a}{2}$. Also, event W is the union of events A and B.

$$\mathsf{Pr}(W) = \mathsf{Pr}(a \leq X \leq b) = \mathsf{Pr}\left(a \leq X \leq \frac{b+a}{2}\right) + \mathsf{Pr}\left(\frac{b+a}{2} \leq X \leq b\right) = \mathsf{Pr}(A) + \mathsf{Pr}(B)$$

$$\mathsf{Pr}(A, B) = 0$$

For discrete random variables, $\mathsf{p}(\cdot)$ is called a *probability mass function* (PMF) and $\mathsf{p}(x) = \mathsf{Pr}(X = x)$.

Cumulative Distribution Functions (CDFs)

When dealing with continuous random variables (or discrete random variables with support over an infinite number of events), the probability of seeing any one particular $x \in \chi$, where χ is the set of all possible values x can take, has measure 0; there are infinitely many points, and we have to assign each a probability such that the probabilities of each of these infinitely many points (hypothetically) sum to one. Instead of focusing on $\mathsf{Pr}(X = x)$, in the continuous case we can consider $\mathsf{Pr}(X \leq x)$. This function is called the *cumulative distribution function* (CDF).

$$\mathsf{F}(x) \triangleq \mathsf{Pr}(X \leq x), x \in \chi$$

The CDF is a monotone increasing function. Taking the derivative of the CDF with respect to x yields the *probability density function* (PDF).

$$\mathsf{p}(x) = \frac{d}{dx}\mathsf{F}(x)$$

The probability that the random variable X is near a specific value x is approximated by:

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$
$$P(x < X \leq x + dx) = F(x + dx) - F(x) \approx \mathsf{p}(x)dx$$

2 Bayesian Statistics: A Very Brief Introduction

Suppose we have some *parameters* θ. These are unknown quantities, such as the average height of all men of age 18 in North Carolina. We have some data D (such as the height of 30,000 North Carolinian men of age 18). How can we describe our uncertainty about θ based on our data? To answer this question, we appeal to Bayes Rule.

$$\mathsf{p}(\theta \mid X) = \frac{\mathsf{p}(X \mid \theta)\mathsf{p}(\theta)}{\mathsf{p}(X)}$$

Let's break down this expression.

- $\mathsf{p}(\theta \mid X)$
  - This is called the *posterior distribution* of θ given X and calculating (or approximating) it is the main goal of Bayesian inference. The posterior expresses our the uncertainty about θ, the parameter(s) we care about, *after* we have seen ("learned from") the data. There is still uncertainty, because while the data told us something about θ, we still don't know everything about it.
- $\mathsf{p}(X \mid \theta)$:
  - This is called the *likelihood*, and is a function of the parameters θ. Basically, this is a way of describing the probability of the data as a function of these unknown parameters. We will discuss this in the third class.
- $\mathsf{p}(\theta)$
  - This is called the *prior distribution*, and it describes our belief about the quantity of interest *before* we see data. What do we think is the average height of 18 year old men in North Carolina, and how certain are we of this belief? Often times we use what are called *conjugate* priors to the likelihood to describe our *a priori* beliefs. More on this soon.
- $\mathsf{p}(X)$
  - This is called the *marginal likelihood* of X, and it is constant with respect to θ. Because of this, we often write the above quantity as $\mathsf{p}(\theta|X) \propto \mathsf{p}(X|\theta)\mathsf{p}(\theta)$.
  - To solve for the marginal likelihood of X:

    $$\mathsf{p}(X) = \int_\Theta \mathsf{p}(X|\theta)\mathsf{p}(\theta)d\theta \tag{1}$$

    Note that this is called the marginal likelihood because we are marginalizing (or averaging) over $\theta \in \Theta$.

2.1 Independence

Many of our calculations will be made simpler when we can assume *independence* between certain random variables. Formally, two events A and B are marginally independent, i.e. $A \perp B$, if their joint density $\mathsf{p}(A, B)$ may be expressed as the product of the marginals:

$$A \perp B \iff \mathsf{p}(A, B) = \mathsf{p}(A)\mathsf{p}(B)$$

Two events A and B are *conditionally independent* given C if $\mathsf{p}(A, B \mid C) = \mathsf{p}(A \mid C)\mathsf{p}(B \mid C)$. This is denoted $(A \perp B) \mid C$.

$$(A \perp B) \mid C \iff \mathsf{p}(A, B \mid C) = \mathsf{p}(A \mid C)\mathsf{p}(B \mid C)$$

Conditional independence allows us to build large, complicated networks of random variables out of simple components for which we can easily perform analysis.

3 Basic Statistics

Expected Value

The expected value of a random variable is also known as the mean and is commonly denoted by μ

- For a discrete random variable x

  $\mathbb{E}[X] = \sum_{x \in X} x\mathsf{p}(x)$

- For a continuous random variable x

  $\mathbb{E}[X] = \int_X x\mathsf{p}(x)dx$

Law of the Unconscious Statistician

Let $h(\cdot)$ be a function of X

- For a discrete random variable x

  $\mathbb{E}[h(X)] = \sum_{x \in X} h(x)\mathsf{p}(x)$

- For a continuous random variable x

  $\mathbb{E}[h(X)] = \int_X h(x)\mathsf{p}(x)dx$

Variance

The variance of a random variable is commonly denoted by σ^2. It is a measure of how dispersed its values are relative to the mean. Small variance indicates that the values are tightly clustered around the mean.

$$\begin{aligned}\mathsf{Var}[X] &\triangleq \mathbb{E}[(x-\mu)^2]\\ &= \int_X (x-\mu)^2 p(x)dx\\ &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2\\ &= \mathbb{E}[X^2] - \mu^2\\ &= \sigma^2.\end{aligned}$$

The *standard deviation* of $X \triangleq \mathsf{sd}[X] \triangleq \sqrt{\mathsf{Var}[X]}$

Since the variance has squared units, which may not always have a natural interpretation, the standard deviation is commonly reported for ease of interpretability. Note that $\mathsf{Var}[X] \geq 0$.

Covariance

The covariance between two random variables measures their degree of linear association. If the covariance is 0, then they are linearly unrelated.

$$\begin{aligned}\mathsf{Cov}[X,Y] &= \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]\\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\end{aligned}$$

Note that the previously given definition of *variance* arises as a special case when $Y = X$.

Correlation

The correlation is another measure of the linear association between two random variables. If the correlation is 0, they are linearly unrelated. If the correlation is ± 1, then there is a perfect positive, or negative linear relationship between them respectively.

$$\mathsf{Cor}(X,Y) \triangleq \mathsf{Cov}(X,Y)/\sqrt{\mathsf{Var}(X)\mathsf{Var}(Y)}$$

Observe that $-1 \leq \mathsf{Cor}(X,Y) \leq 1$.

4 Some Common Discrete Distributions

The notation $X \sim$ Distribution(parameters) indicates that X is a random variable that follows the named distribution with given parameters.

Bernouli Distribution

Notation:	$X \sim \mathsf{Ber}(\theta)$
	where $0 \leq \theta \leq 1$
Support:	$\{0, 1\}$
PMF:	$\theta^x(1-\theta)^{1-x}$
Expectation:	θ
Variance:	$\theta(1-\theta)$

Example: Let X denote the result of a single coin flip, where tails correspond to $X = 0$, heads corresponds to $X = 1$, and θ is the probability of heads.

Bionomial Distribution

Notation:	$X \sim \mathsf{Bin}(n, \theta)$
	where $n > 0$ and $0 \leq \theta \leq 1$
Support:	$\{0, 1, \ldots, n\}$
PMF:	$\binom{n}{x}\theta^x(1-\theta)^{n-x}$
Expectation:	$n\theta$
Variance:	$n\theta(1-\theta)$

Example: Let X denote the number of heads in n flips of the same coin where θ is the probability of heads on each individual flip. Recall that $\binom{n}{x} = n!/(x!(n-x)!)$, where ! represents the factorial function. This represents the number of sequences of n coin flips that have exactly x heads.

Exchangeability

Recall that the coin flips are exchangeable in this distribution. The probability of any sequence of flips depends *only* on the number of heads, the number of tails, and the probability of heads on each flip.
The order in which we observed the coin flips does not matter.

Multinomial Distribution

Notation:	$X \sim \mathsf{Mult}(n, \theta)$
	where $n > 0$, $\theta = [\theta_1, \ldots, \theta_k]$ and $\sum \theta_j = 1$
Support:	$\{x_1, \ldots, x_k \mid \sum x_j = n, x_j \geq 0\}$
PMF:	$\binom{n}{x_1,\ldots,x_k}\prod \theta_j^{x_j}$
$\mathbb{E}[X] =$	$n\theta$
$\mathsf{Var}[X_j] =$	$n\theta_j(1-\theta_j)$
$\mathsf{Cov}[X_i, X_j] =$	$-n\theta_i\theta_j$

The multinomial generalizes the binomial to arbitrary collections of categorical variables.

Example: Let X denote the vector of the number of times each side of a k sided die has landed face up in n tosses of the die, and let θ_j be the probability that the number 'j' is rolled on each toss of the die. For example, x_1 represents the number of times a ‘1’ was rolled, x_2 represents the number of times a ‘2’ was rolled, etc.

In this example the categories are $1, \ldots, k$, but in general they can be completely arbitrary. There is no need to make them ordered, or even numeric.

Poisson Distribution

Notation:	$X \sim \mathsf{Pois}(\lambda)$ where $\lambda > 0$
Support:	$\{0, 1, 2, 3, \ldots\}$ (all non-negative integers)
PMF:	$e^{-\lambda}\lambda^x/x!$
Expectation:	λ
Variance:	λ

Typical Examples Include: Data are based on counts over a specific location or time. For example:

- the number of cars driving down a street every hour
- the number of customers in line
- the number of mutations along a stretch of genome.

Continuous Distributions

Gaussian (Normal) Distribution

Notation:	$X \sim \mathcal{N}(\mu, \sigma^2)$ where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$
Support:	$\mathbb{R}$ (the real numbers)
PDF:	$\sqrt{\frac{1}{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$
Expectation:	μ
Variance:	σ^2

$1/\theta^2$ is often called the *precision*. A Gaussian distribution with a mean of 0 and a standard deviation of 1 is known as the *standard normal* distribution.

Due to the Central Limit Theorem, this is probably the most common distribution. For example, let X denote the heights of the males in our class, then X is approximately normally distributed.

Multivariate Gaussian Distribution

Notation:	$X \sim \mathcal{N}(\mu, \Sigma)$ where $\mu \in \mathbb{R}^k$ and Σ is $k \times k$ symmetric, positive definite
Support:	$\mathbb{R}^k$
PDF:	$\left(\frac{1}{2\pi}\right)^{k/2} \lvert\Sigma\rvert^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$
Expectation:	μ
Variance:	Σ

The diagonal elements of Σ are the variances of each X_j, and, in general, the i, j entry of Σ is the covariance between X_i and X_j. Positive definite means that we have $\boldsymbol{v}\boldsymbol{\Sigma}\boldsymbol{v}^T > 0$, for all $\boldsymbol{v} \in \mathbb{R}^k$

Additional Material

Joe Blitzstein's Statistics 101

Video lectures of a full introductory course on statistics by Joe Blitzstein, Harvard University, are freely available on iTunes U (`https://itunes.apple.com/us/course/statistics-110-probability/id502492375`). Fundamental concepts such as conditioning, independence, moments, as well as the most important distributions and their relationships to each other are all introduced in a very intuitive way. The course has very few prerequisites, and the "Final Review" PDF is a compact summary of nearly everything taught in the course.

Mathematicalmonk's Machine Learning, video series on Youtube

In his Youtube channel (`http://www.youtube.com/user/mathematicalmonk`), the author provides introductory videos to many relevant concepts, e.g. the multivariate gaussian distribution.