

CSE 6740 Lecture 2

How Do I Learn a Simple Model? (Probability and inference)

Alexander Gray

agray@cc.gatech.edu

Georgia Institute of Technology

Today

1. Probability and random variables (*What is “data”? What is a “model”?*)
2. Asymptotics and point estimation (*What is “estimation/learning”?*)
3. Confidence intervals (*How good is the estimation/learning?*)

Mathematical Notation

Pay attention to my comments about the notation for various things. Statistical notation is highly variable and ambiguous and can make things confusing, all on its own.

Probability and random variables

What is “data”? What is a “model”?

Samples Spaces and Events

If we toss a coin twice then the *sample space*, or set of all possible outcomes or realizations ω , is

$$\Omega = \{HH, HT, TH, TT\}.$$

An *event* is a subset of this set; for example the event that the first toss is heads is $A = \{HH, HT\}$.

Probability

We'll assign a real number $\mathbb{P}(A)$ to each event A , called the *probability* of A . To qualify as a probability, \mathbb{P} must satisfy three axioms:

1. $\mathbb{P}(A) \geq 0$ for every A
2. $\mathbb{P}(\Omega) = 1$
3. If A_1, A_2, \dots are disjoint then

$$\mathbb{P} \left(\bigcup_{i=1}^{\infty} A_i \right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i). \quad (1)$$

Note that frequentists and Bayesians agree on these.

Random Variables

A *random variable* is a mapping, or function

$$X : \Omega \rightarrow \mathbb{R} \quad (2)$$

that assigns a real number $X(\omega)$ to each outcome ω .

For example, if $\Omega = \{(x, y) : x^2 + y^2 \leq 1\}$ and our outcomes are samples (x, y) from the unit disk, then these are some random variables: $X(\omega) = x$, $Y(\omega) = y$, $Z(\omega) = x + y$.

Data and Statistics

The *data* are specific values of random variables.

A *statistic* is just any function of the data/random variables.
Any function of a random variable is itself a random variable.

Distribution Functions

Suppose X is a random variable, x a specific value of it (data).

Cumulative distribution function (CDF): the function $F : \mathbb{R} \rightarrow [0, 1]$ (sometimes F_X) defined by $F(x) = \mathbb{P}(X \leq x)$.

X is *discrete* if it takes countably many values $\{x_1, x_2, \dots\}$.

Probability (mass) function for X : $f(x) = \mathbb{P}(X = x)$.

Distribution Functions

X is *continuous* if there exists a function f such that $f(x) \geq 0$ for all x , $\int_{-\infty}^{\infty} f(x)dx = 1$ and for every $a \leq b$,

$$\mathbb{P}(a < X < b) = \int_a^b f(x)dx. \quad (3)$$

f is the *probability density function* (PDF).

We have that $F(x) = \int_{-\infty}^x f(t)dt$ and $f(x) = F'(x)$ wherever F is differentiable.

Discrete Distributions

Some examples of discrete distributions:

X is the outcome of a coin flip. $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$ for some $p \in [0, 1]$. We say $X \sim \text{Bernoulli}(p)$. $f(x) = p^x(1 - p)^{1-x}$ for $x \in \{0, 1\}$.

Binomial: the distribution of the *number* of outcomes (of say, heads) of a coin flip.

Continuous Distributions

Some examples of continuous distributions:

Uniform: $X \sim \text{Uniform}(a, b)$ if $f(x) = 1/(b - a)$ for $x \in [a, b]$, 0 otherwise.

Gaussian: $X \sim \mathcal{N}(\mu, \sigma^2)$ if $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\}$ for $\mu \in \mathbb{R}, \sigma > 0$. We call its PDF $\phi(x)$ and its CDF $\Phi(x)$.

Standard Normal Distribution

We say that a random variable has a *standard Normal* distribution if $\mu = 0$ and $\sigma = 1$, and we denote it by Z .

If $X \sim \mathcal{N}(\mu, \sigma^2)$ then $Z = (X - \mu)/\sigma \sim \mathcal{N}(0, 1)$.

If $Z \sim \mathcal{N}(0, 1)$ then $X = \mu + \sigma Z \sim \mathcal{N}(\mu, \sigma^2)$.

Multivariate Distributions

Can define a distribution over a vector of random variables. We say this is a *multivariate* distribution.

Our dataset generally consists of samples from a multivariate distribution. Each of the columns is a random variable. We can also consider the whole vector of random variables as a random variable.

Marginal Distributions

Discrete case: Suppose (X, Y) have joint distribution with mass function f . The *marginal mass function* for X is defined by

$$f(x) = \mathbb{P}(X = x) = \sum_y \mathbb{P}(X = x, Y = y) = \sum_y f(x, y). \quad (4)$$

Continuous case: *Marginal density* $f(x) = \int f(x, y)dy$.

Conditional Distributions

For X and Y discrete, the distribution of X given that we have observed $Y = y$ is the *conditional probability function*

$$f(x|y) = \mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{f(x, y)}{f(y)} \quad (5)$$

if $f(y) > 0$. For X and Y continuous, the *conditional PDF* is

$$f(x|y) = \frac{f(x, y)}{f(y)} \text{ if } f(y) > 0. \text{ Then}$$

$$\mathbb{P}(X \in A|Y = y) = \int_A f(x|y) dx.$$

Note that $f(x, y) = f(x|y)f(y) = f(y|x)f(x)$.

Bayes' Rule

If X takes values x_1, \dots, x_n and y is a value of Y ,

$$f(y) = \sum_j f(y|x_j) f(x_j) \quad (6)$$

and so

$$f(x_i|y) = \frac{f(y|x_i) f(x_i)}{\sum_j f(y|x_j) f(x_j)}. \quad (7)$$

The latter is called *Bayes' rule*.

Continuous case:

$$f(x|y) = \frac{f(y|x) f(x)}{\int f(y|x) f(x) dx}. \quad (8)$$

Independence

X and Y are *independent* if

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y) \quad (9)$$

or $f(x, y) = f(x)f(y)$ for all values x and y .

Independent identically distributed (IID) random variables are drawn from the same distribution and independent.

Expectation

The *expected value*, or *mean*, or *first moment* of X is

$$\mathbb{E}(X) = \mathbb{E}X = \mu = \int x f(x) dx. \quad (10)$$

Note that in the discrete case this means $\sum_x x f(x)$.

$$\mathbb{E} \left(\sum_i a_i X_i \right) = \sum_i a_i \mathbb{E}(X_i) \quad (11)$$

for constants a_1, a_2, \dots, a_n . If the X_i are independent,

$$\mathbb{E} \left(\prod_i a_i X_i \right) = \prod_i a_i \mathbb{E}(X_i). \quad (12)$$

Variance

The k^{th} moment of X is defined to be $\mathbb{E}(X^k)$ assuming that $\mathbb{E}(X^k) < \infty$.

If X has mean μ , the *variance* of X is

$$\sigma^2 = \mathbb{V}(X) = \mathbb{V}X = \mathbb{E}(X - \mu)^2 = \int (x - \mu)^2 f(x) \quad (13)$$

and $\sigma = \text{sd}(X) = \sqrt{\mathbb{V}(X)}$.

Sample Statistics

If X_1, \dots, X_n are random variables then the *sample mean* is

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i \quad (14)$$

and the *sample variance* is

$$S^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2. \quad (15)$$

If X_1, \dots, X_n are IID, then

$$\mathbb{E}(\bar{X}) = \mathbb{E}(X_i) = \mu, \quad \mathbb{V}(\bar{X}) = \sigma^2/N, \quad \mathbb{E}(S^2) = \sigma^2. \quad (16)$$

Asymptotic theory and point estimation

What is “estimation/learning”? What happens as you get more data? Why is the sample mean a good estimator?

Convergence

Suppose X_1, X_2, \dots is a sequence of random variables and X is another random variable. F_N is the CDF of X_N and F is the CDF of X .

X_N converges in probability to X , $X_N \xrightarrow{p} X$, if for every $\epsilon > 0$,

$$\mathbb{P}(|X_N - X| > \epsilon) \rightarrow 0 \quad (17)$$

as $N \rightarrow \infty$.

Convergence

X_N converges in distribution to X , $X_N \rightsquigarrow X$, if

$$\lim_{N \rightarrow \infty} F_N(t) = F(t) \quad (18)$$

at all t for which F is continuous.

Convergence

These are ordered in strength:

$$X_N \xrightarrow{p} X \Rightarrow X_N \rightsquigarrow X \quad (19)$$

Special case: if $\mathbb{P}(X = c) = 1$ for some $c \in \mathbb{R}$,

$X_N \rightsquigarrow X \Rightarrow X_N \xrightarrow{p} X$. But in general none of the reverse implications hold.