ISTA 421 + INFO 521
Introduction to Machine Learning

Lecture 7:
More Probability

Clay Morrison
claytonm@email.arizona.edu

13 September 2017

1

Next Topics

- Expectation and Random Vectors, Covariance
- Discrete Probability
  - Probability Mass Function (pmf)
  - Example discrete distributions (Bernoulli, Binomial)
- Continuous probability
  - Probability Density Function (pdf)
  - Gaussian Distribution
- Return to the Linear Model, with Noise!
  - Likelihood Function
  - Maximum Likelihood Estimation

2

Expectation

The **expected value** of a function of a random variable *X* that is distributed according to P(X) is:

$$\mathbf{E}_{P(x)}\{f(X)\} = \sum_x f(x)P(x)$$

The expected value of a (function of a) random variable is the **weighted (by probability) average** of all possible values of that variable (through that function).

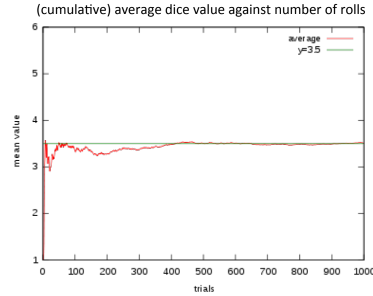
The expected value of the random variable *X* itself: the **mean**

$$\mathbf{E}_{P(x)}\{X\} = \sum_x xP(x)$$

What is the relationship of the *arithmetic mean* to the expected value?

$$= \frac{1}{N}\sum_{i=1}^{N} x_i$$

(cumulative) average dice value against number of rolls



Expectation

$$\mathbf{E}_{P(x)}\{f(X)\} = \sum_x f(x)P(x)$$

The expectation of the value of *X* if *X* is a fair die:

$$\mathbf{E}_{P(x)}\{X\} = \sum_x x\frac{1}{6} = \frac{1}{6} + \frac{2}{6} + \ldots + \frac{6}{6} = \frac{21}{6} = 3.5$$

SISTA 4

Expectation

$$\mathbf{E}_{P(x)}\{f(X)\} = \sum_x f(x)P(x)$$

The expectation of the value of *X* if *X* is a fair die:

$$(\mathbf{E}_{P(x)}\{X\})^2 = (\sum_x x\frac{1}{6} = \frac{1}{6} + \frac{2}{6} + \ldots + \frac{6}{6} = \frac{21}{6})^2 = (3.5)^2 = 12.25$$

$$\mathbf{E}_{P(x)}\{X^2\} = \sum_x x^2\frac{1}{6} = \frac{1}{6} + \frac{4}{6} + \ldots + \frac{36}{6} = \frac{91}{6} \approx 15.17$$

$$12.25 \neq 15.17$$

$$(\mathbf{E}_{P(x)}\{X\})^2 \neq \mathbf{E}_{P(x)}\{X^2\}$$

5

Expectation

$$\mathbf{E}_{P(x)}\{f(X)\} = \sum_x f(x)P(x)$$

The expectation of the value of *X* if *X* is a fair die:

$$(\mathbf{E}_{P(x)}\{X\})^2 = (\sum_x x\frac{1}{6} = \frac{1}{6} + \frac{2}{6} + \ldots + \frac{6}{6} = \frac{21}{6})^2 = (3.5)^2 = 12.25$$

$$\mathbf{E}_{P(x)}\{X^2\} = \sum_x x^2\frac{1}{6} = \frac{1}{6} + \frac{4}{6} + \ldots + \frac{36}{6} = \frac{91}{6} \approx 15.17$$

$$12.25 \neq 15.17$$

$$(\mathbf{E}_{P(x)}\{X\})^2 \neq \mathbf{E}_{P(x)}\{X^2\}$$

More precisely:
Jensen's Inequality

$$(\mathbf{E}_{P(x)}\{X\})^2 \leq \mathbf{E}_{P(x)}\{X^2\}$$

6

Expectation

$$\mathbf{E}_{P(x)}\left\{f(X)\right\} = \sum_x f(x)P(x)$$

In ***general***: the expected value of a function of *X* is ***not equal*** to the function evaluated at the expected value of *X*!

usually

$$f(\mathbf{E}_{P(x)}\{X\}) \neq \mathbf{E}_P(x)\{f(X)\}$$

BUT! These cases *do* hold:

$$f(X) = a \quad : \quad \mathbf{E}_{P(x)}\{f(X)\} = a$$
$$f(X) = aX \quad : \quad \mathbf{E}_{P(x)}\{f(aX)\} = a\mathbf{E}_{P(x)}\{f(X)\}$$
$$\mathbf{E}_{P(x)}\{f(X) + g(X)\} = \mathbf{E}_{P(x)}\{f(X)\} + \mathbf{E}_{P(x)}\{g(X)\}$$

SISTA 7

Expectation: Variance

$$\mathbf{E}_{P(x)}\left\{f(X)\right\} = \sum_x f(x)P(x)$$

Variance:

$$\text{var}\{X\} = \mathbf{E}_{P(x)}\left\{(X - \mathbf{E}_{P(x)}\left\{x\right\})^2\right\}$$

$$\begin{aligned}
\text{var}\{X\} &= \mathbf{E}_{P(x)}\left\{(X - \mathbf{E}_{P(x)}\left\{X\right\})^2\right\} \\
&= \mathbf{E}_{P(x)}\left\{X^2 - 2X\mathbf{E}_{P(x)}\left\{X\right\} + \mathbf{E}_{P(x)}\left\{X\right\}^2\right\} \\
&= \mathbf{E}_{P(x)}\left\{X^2\right\} - 2\mathbf{E}_{P(x)}\left\{X\right\}\mathbf{E}_{P(x)}\left\{X\right\} + \mathbf{E}_{P(x)}\left\{X\right\}^2
\end{aligned}$$

$$\text{var}\{X\} = \mathbf{E}_{P(x)}\left\{X^2\right\} - \mathbf{E}_{P(x)}\left\{X\right\}^2$$

SISTA 8

Vector Random Variables

Vector random variables!

$p(\mathbf{x}) = p(x_1, x_2, \ldots, x_N) = P(X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N)$

Mean: $\mathbf{E}_{P(\mathbf{x})}\{\mathbf{x}\} = \sum_{\mathbf{x}} \mathbf{x}P(\mathbf{x})$

Very similar to ***scalar*** version:

$$\mathbf{E}_{P(x)}\{X\} = \sum_{x} xP(x)$$

When we move to vector random variables and consider their "variance", the scalar version of variance needs to be extended…

Scalar variance:

$$\text{var}\{X\} = \mathbf{E}_{P(x)}\{(X - \mathbf{E}_{P(x)}\{x\})^2\}$$

$$\text{var}\{X\} = \mathbf{E}_{P(x)}\{X^2\} - \mathbf{E}_{P(x)}\{X\}^2$$

$$var\{X\} = \mathbf{E}_{P(x)}\{(X - \mathbf{E}_{P(x)}\{x\})(X - \mathbf{E}_{P(x)}\{x\})\}$$

The scalar "summation" form of variance:

$$\begin{aligned} var(X) &= \sum_x (x - \mu_X)^2 \\ &= \sum_x (x - \mu_X)(x - \mu_X) \end{aligned}$$

This is comparing a variable to itself

When we want to calculate how one random variable (co)varies with another, Then we are interested in the **co**variance:

$$cov(X, Y) = \mathbf{E}_{p(x,y)}\left\{\left(x - \mathbf{E}_{p(x)}\{x\}\right)\left(y - \mathbf{E}_{p(y)}\{y\}\right)\right\}$$

SISTA 9

(Co)variance of a Random Vector

- Covariance

$$cov(X, Y) = \mathbf{E}_{p(x,y)}\left\{\left(x - \mathbf{E}_{p(x)}\{x\}\right)\left(y - \mathbf{E}_{p(y)}\{y\}\right)\right\}$$

- Now, if we want to take the "variance" of a random vector, which is essentially a compact representation of a ***joint distribution***, then we need to keep track of all of the pair-wise **co**variances of each of the random vector components, and we do this in the ***covariance matrix***:

$$\Sigma = \begin{bmatrix} \mathrm{E}[(X_1-\mu_1)(X_1-\mu_1)] & \mathrm{E}[(X_1-\mu_1)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_1-\mu_1)(X_n-\mu_n)] \\ \mathrm{E}[(X_2-\mu_2)(X_1-\mu_1)] & \mathrm{E}[(X_2-\mu_2)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_2-\mu_2)(X_n-\mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[(X_n-\mu_n)(X_1-\mu_1)] & \mathrm{E}[(X_n-\mu_n)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_n-\mu_n)(X_n-\mu_n)] \end{bmatrix}.$$

cov{*x*} is shorthand for cov(*x*,*x*)
When x is a random vector, **x**, then this is a matrix, Σ

SISTA 10

Vector Random Variables

Vector random variables!

$p(\mathbf{x}) = p(x_1, x_2, \ldots, x_N) = P(X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N)$

Mean: $\mathbf{E}_{P(\mathbf{x})}\{\mathbf{x}\} = \sum_{\mathbf{x}} \mathbf{x} P(\mathbf{x})$

Very similar to ***scalar*** version:

$$\mathbf{E}_{P(x)}\{X\} = \sum_{x} x P(x)$$

Covariance:

$$\text{cov}\{\mathbf{x}\} = \mathbf{E}_{P(\mathbf{x})}\left\{\left(\mathbf{x} - \mathbf{E}_{P(\mathbf{x})}\{\mathbf{x}\}\right)\left(\mathbf{x} - \mathbf{E}_{P(\mathbf{x})}\{\mathbf{x}\}\right)^{\mathsf{T}}\right\}$$

$$\begin{aligned}\text{cov}\{\mathbf{x}\} &= \mathbf{E}_{P(\mathbf{x})}\left\{\left(\mathbf{x} - \mathbf{E}_{P(\mathbf{x})}\{\mathbf{x}\}\right)\left(\mathbf{x} - \mathbf{E}_{P(\mathbf{x})}\{\mathbf{x}\}\right)^{\mathsf{T}}\right\} \\ &= \mathbf{E}_{P(\mathbf{x})}\left\{\mathbf{x}\mathbf{x}^{\mathsf{T}} - 2\mathbf{x}\mathbf{E}_{P(\mathbf{x})}\{\mathbf{x}\}^{\mathsf{T}} + \mathbf{E}_{P(\mathbf{x})}\{\mathbf{x}\}\mathbf{E}_{P(\mathbf{x})}\{\mathbf{x}\}^{\mathsf{T}}\right\}\end{aligned}$$

$$\text{cov}\{\mathbf{x}\} = \mathbf{E}_{P(\mathbf{x})}\left\{\mathbf{x}\mathbf{x}^{\mathsf{T}}\right\} - \mathbf{E}_{P(\mathbf{x})}\{\mathbf{x}\}\mathbf{E}_{P(\mathbf{x})}\{\mathbf{x}\}^{\mathsf{T}}$$

SISTA 11

Discrete Distributions:

Probability *Mass* Functions (pmf)

SISTA 12

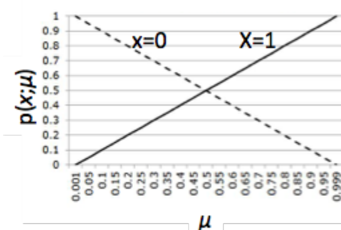
See FCML Ch 2.3, PRML Ch2 (posted on D2L).

Bernoulli Distribution

$$x \in \{0, 1\} \quad \text{(e.g., 1 is "heads" and 0 is "tails")}$$

$$p(x = 1|\mu) = \mu \quad \text{and} \quad p(x = 0|\mu) = 1 - \mu$$

$$p(x|\mu) = \begin{cases} \mu & \text{if } x = 1 \\ 1 - \mu & \text{if } x = 0 \end{cases}$$



13

See FCML Ch 2.3, PRML Ch2 (posted on D2L).

Bernoulli Distribution

$$x \in \{0, 1\} \quad \text{(e.g., 1 is "heads" and 0 is "tails")}$$

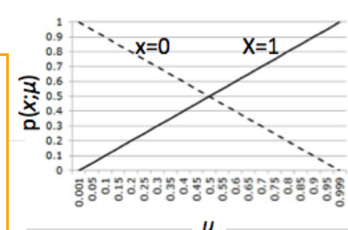
$$p(x = 1|\mu) = \mu \quad \text{and} \quad p(x = 0|\mu) = 1 - \mu$$

$$Bern(x|\mu) = \mu^x (1 - \mu)^{(1-x)}$$

Study this trick!

x is an *indicator variable* which is constrained to be "1" for exactly one value, and "0" for the rest.

While it looks like it makes things complicated, the "if" in the previous is awkward in formulas.



14

See FCML Ch 2.3, PRML Ch2 (posted on D2L).

Binomial Distribution

Probability distribution for getting *m* "heads" out of *N* tosses.

$$Bin(m|N,\mu) = \underbrace{\binom{N}{m}}_{\text{Number of ways to get } m \text{ heads in } N \text{ tosses}} \cdot \underbrace{\mu^m (1-\mu)^{(N-m)}}_{\text{Probability of \textbf{each} way to get } m \text{ heads in } N \text{ tosses}}$$

Example event
N=3, *m*=2
HHT
HTH
THH

where $\binom{N}{m} = \dfrac{N!}{(N-m)!m!}$

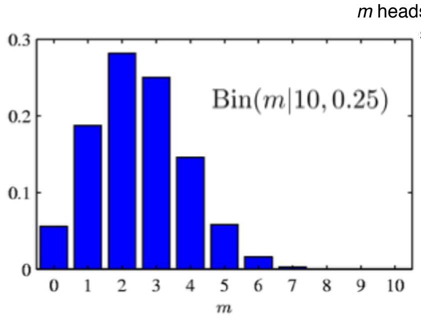
15

See FCML Ch 2.3, PRML Ch2 (posted on D2L).

Binomial Distribution

Probability distribution for getting *m* "heads" out of *N* tosses.

$$Bin(m|N,\mu) = \underbrace{\binom{N}{m}}_{\text{Number of ways to get } m \text{ heads in s}} \cdot \underbrace{\mu^m (1-\mu)^{(N-m)}}_{\text{Probability of \textbf{each} way to get } m \text{ heads in } N \text{ tosses}}$$



where $\binom{N}{m} = \dfrac{N!}{(N-m)!m!}$

16