ISTA 421/521
Introduction to Machine Learning

Lecture 6:
Probability Review,
Expectation,
Discrete Prob. Distributions

Clay Morrison

clayton@sista.arizona.edu

Gould-Simpson 819

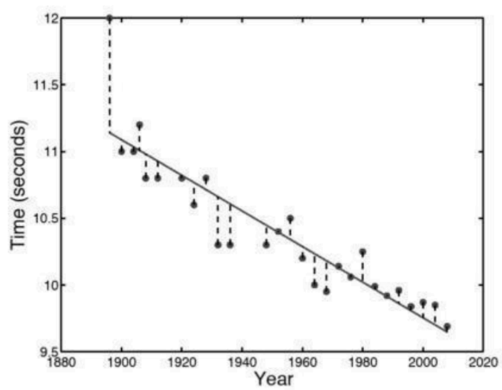
Phone 621-6609

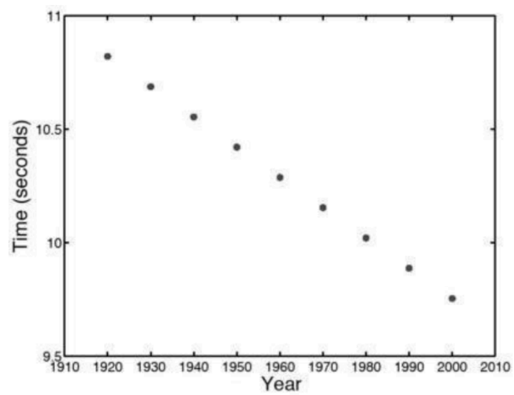
11 September 2014

Next Topics

- Probability Basics
- Expectation and Random Vectors
- Discrete Probability
- Example discrete distributions
- Continuous probability
- Gaussian Distribution
- Maximum Likelihood Estimation

Think Generatively

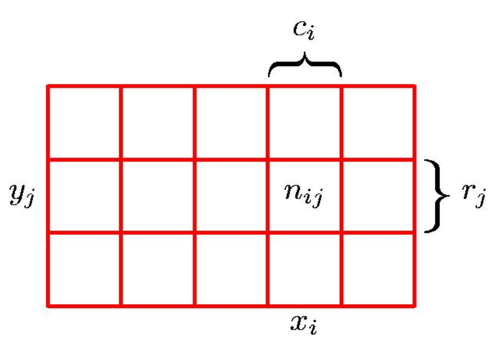




Data generated from linear model

Want degree of confidence in parameter values and predictions

Probability Theory



•Marginal Probability

$$p(X = x_i) = \frac{c_i}{N}.$$

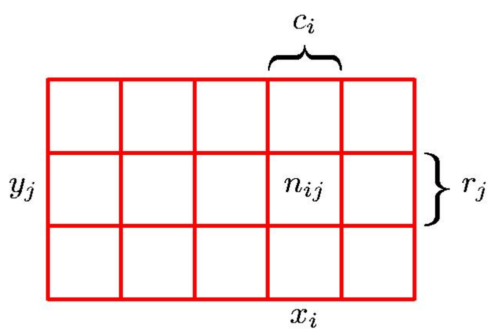
Joint Probability

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

•Conditional Probability

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

Probability Theory



• Sum Rule

$$p(X = x_i) = \frac{c_i}{N} = \frac{1}{N} \sum_{j=1}^{L} n_{ij}$$

$$= \sum_{j=1}^{L} p(X = x_i, Y = y_j)$$

Product Rule

$$\begin{aligned} p(X = x_i, Y = y_j) &= \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N} \\ &= p(Y = y_j | X = x_i) p(X = x_i) \end{aligned}$$

The Rules of Probability

- Sum Rule $\qquad p(X) = \sum_{Y} p(X, Y)$
- Product Rule $\qquad p(X, Y) = p(Y|X)p(X)$

Bayes' Theorem

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

$$p(X) = \sum_Y p(X|Y)p(Y)$$

posterior ∝ likelihood × prior

Expectation

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

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

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

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

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:

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

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

$12.25 \neq 15.16$

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

Expectation

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

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

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

$$f(\mathbf{E}_{P(x)}\{X\}) \overset{\textit{usually}}{\neq} \mathbf{E}_{P(x)}\{f(X)\}$$

Special cases:

$f(X) = aX$

$f(X) = a$

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

Expectation: Variance

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

Variance:

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

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

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

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:

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

Covariance:

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

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

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

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:

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

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

Scalar variance:

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

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

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

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 like 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:

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

(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***:

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

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

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

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

Discrete Distributions

- The functions that characterize the discrete random variable are often referred to as *probability mass functions* (pmf)

- Bernoulli
- Binomial
- Multinomial

Bernoulli Distribution

- Coin flipping: heads=1, tails=0

$$p(x=1|\mu) = \mu$$

- Bernoulli Distribution

$$\begin{aligned} \mathrm{Bern}(x|\mu) &= \mu^x(1-\mu)^{1-x} \\ \mathbb{E}[x] &= \mu \\ \mathrm{var}[x] &= \mu(1-\mu) \end{aligned}$$

Binomial Distribution

- N coin flips:

$$p(m \text{ heads}|N, \mu)$$

- Binomial Distribution

$$\mathrm{Bin}(m|N,\mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}$$

$$\mathbb{E}[m] \equiv \sum_{m=0}^{N} m\mathrm{Bin}(m|N,\mu) = N\mu$$

$$\mathrm{var}[m] \equiv \sum_{m=0}^{N} (m - \mathbb{E}[m])^2 \,\mathrm{Bin}(m|N,\mu) = N\mu(1-\mu)$$

Binomial Distribution