CS 591, Lecture 2
Data Analytics: Theory and Applications
Boston University

Charalampos E. Tsourakakis

January 25rd, 2017

Probability Theory

The theory of probability is a system for making better guesses.
http://www.feynmanlectures.caltech.edu/I_06.html

By the “probability” of a particular outcome of an observation we mean our estimate for the most likely fraction of a number of repeated observations that will yield that particular outcome.

http://www.feynmanlectures.caltech.edu/I_06.html

$$\boxed{p(A) = \frac{N_A}{N}}$$

Inclusion Exclusion theorem

Theorem Suppose $n \in \mathbb{N}$ and A_i is a finite set for $1 \leq i \leq n$. It follows that

$$\left| \bigcup_{1 \leq i \leq n} A_i \right| = \sum_{1 \leq i_1 \leq n} |A_{i_1}| - \sum_{1 \leq i_1 \leq i_2 \leq n} |A_{i_1} \cap A_{i_2}|$$
$$+ \sum_{1 \leq i_1 \leq i_2 \leq i_3 \leq n} |A_{i_1} \cap A_{i_2} \cap A_{i_3}| - \ldots + (-1)^{n+1} \left| \bigcap_{i=1}^{n} A_i \right|$$

Application (aka matching hat problem): Deal two packs of shuffled cards simultaneously. What is the probability that no pair of identical cards will be exposed simultaneously?

Inclusion Exclusion theorem

- Fix the first pack
- Let A_i be the set of all possible arrangements of the second pack which match the card in position i of the first pack.
- $X = \cup_i A_i$

Details on whiteboard.

$$
\begin{aligned}
|X|/52! &= (52!)^{-1}\left(\binom{52}{1}51! - \binom{52}{2}50! + \binom{52}{3}49! - .. - \binom{52}{52}0!\right) \\
&= 1 - 1/2! + 1/3! - \ldots - 1/52! \\
&\approx 1 - \Big(\sum_{i=0}^{\infty}(-1)^i/i!\Big) \\
&= 1 - 1/e.
\end{aligned}
$$

Thus the desired probability is $1/e$ as $n \to +\infty$.

Fundamental Rules

$$\mathbf{Pr}\left[X \vee Y\right] = \mathbf{Pr}\left[X\right] + \mathbf{Pr}\left[Y\right] - \mathbf{Pr}\left[X \wedge Y\right] \tag{1}$$

$$\mathbf{Pr}\left[X\right] = \sum_y \mathbf{Pr}\left[X, Y = y\right] = \sum_y \mathbf{Pr}\left[X | Y = y\right] \mathbf{Pr}\left[Y = y\right] \tag{2}$$

Sum Rule

$$\mathbf{Pr}\left[X, Y\right] = \mathbf{Pr}\left[X \wedge Y\right] = \mathbf{Pr}\left[X\right] \mathbf{Pr}\left[Y | X\right] = \mathbf{Pr}\left[Y\right] \mathbf{Pr}\left[X | Y\right] \tag{3}$$

Product Rule

Fundamental Rules

By applying the product rule multiple times we obtain the chain rule:

$$\mathbf{Pr}\left[X_1, X_2, \ldots, X_n\right] = \mathbf{Pr}\left[X_1\right]\mathbf{Pr}\left[X_2, \ldots, X_n | X_1\right] = .. =$$

$$\mathbf{Pr}\left[X_1\right]\mathbf{Pr}\left[X_2 | X_1\right]\mathbf{Pr}\left[X_3 | X_2, X_1\right] \ldots \mathbf{Pr}\left[X_n | X_1, \ldots, X_{n-1}\right] \quad (4)$$

Chain Rule

$$\mathbf{Pr}\left[X | Y\right] = \frac{\mathbf{Pr}\left[X \wedge Y\right]}{\mathbf{Pr}\left[Y\right]} \quad (5)$$

Conditional probability

Reminder: Bayes' rule

Bayes' rule is a direct application of conditional probabilities.

$$\mathbf{Pr}\left[H|D\right] = \frac{\mathbf{Pr}\left[D|H\right]\mathbf{Pr}\left[H\right]}{\mathbf{Pr}\left[D\right]}, \text{and } \mathbf{Pr}\left[D\right] > 0, \text{or } \ldots$$

posterior $\propto$ likelihood $\times$ prior.

Independence and Conditional Independence

- We say X and Y are *unconditionally independent* or *marginally independent*, or just *independent* if

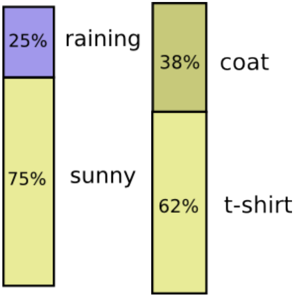
$$\mathbf{Pr}\left[X|Y\right] = \mathbf{Pr}\left[X\right], \mathbf{Pr}\left[Y|X\right] = \mathbf{Pr}\left[Y\right]$$

- As a result

$$\mathbf{Pr}\left[X, Y\right] = \mathbf{Pr}\left[X\right]\mathbf{Pr}\left[Y\right].$$

- **Notation:** $X \perp\!\!\!\perp Y$

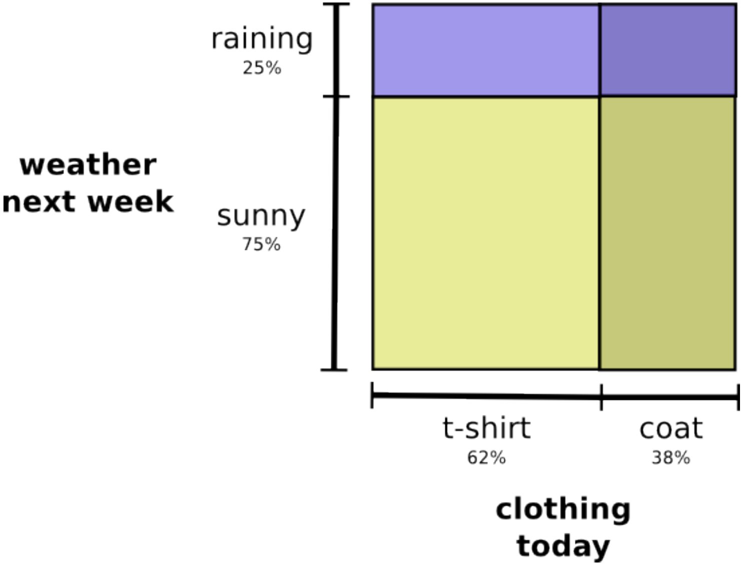
Independence and Conditional Independence



Source: `http://colah.github.io/posts/2015-09-Visual-Information/`

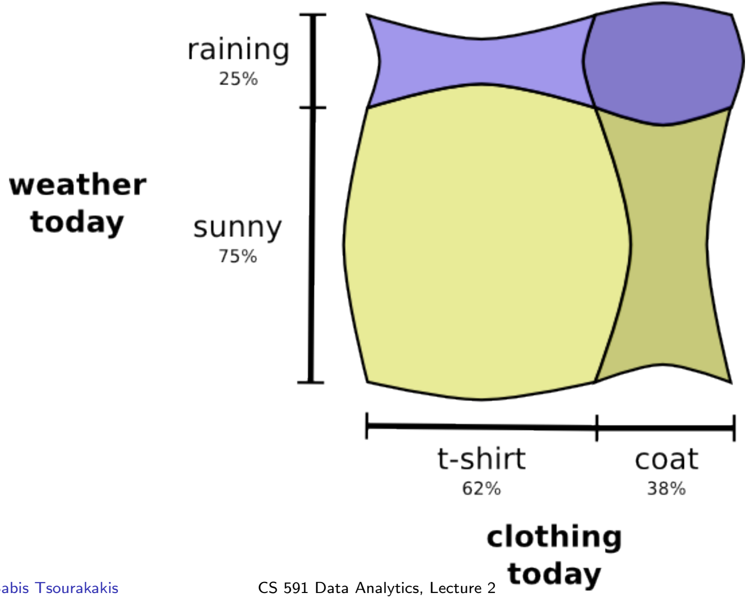
Independence and Conditional Independence

Independence visualized



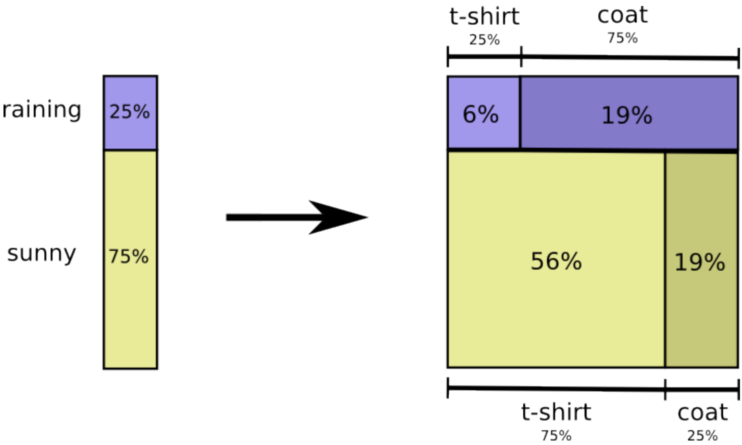
Independence and Conditional Independence

Closer to reality



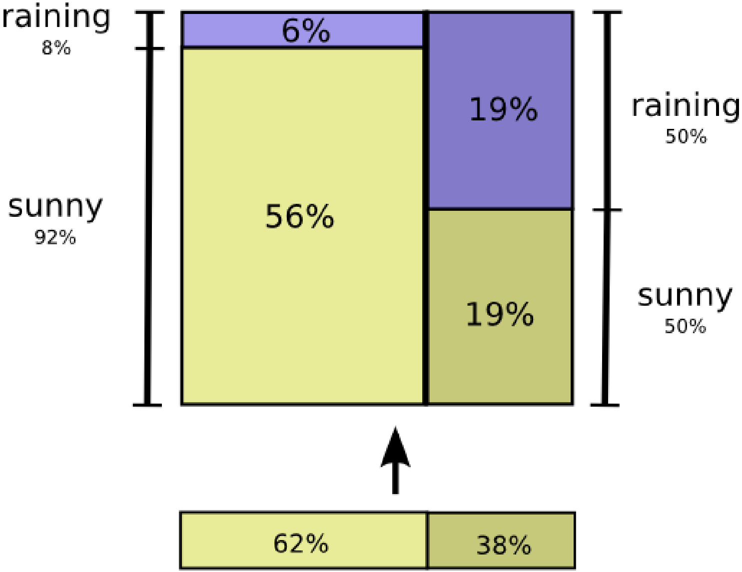
Independence and Conditional Independence

Closer to reality ...



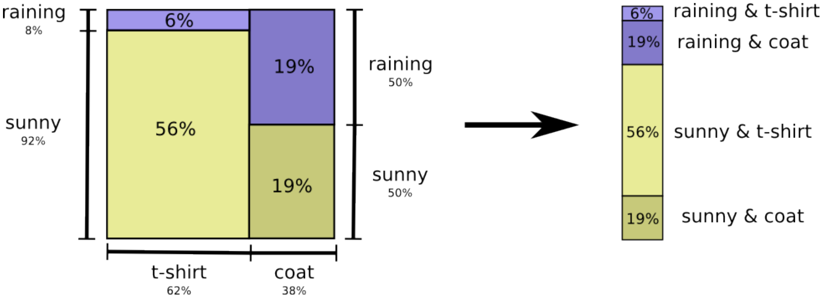
Independence and Conditional Independence

... or alternatively ...



Independence and Conditional Independence

... or alternatively ...



Independence and Conditional Independence

- We say X and Y are *conditionally independent* given Z if

$$\mathbf{Pr}[X, Y|Z] = \mathbf{Pr}[X|Z]\mathbf{Pr}[Y|Z].$$

- Joint distribution factorizes as

$$\mathbf{Pr}[X, Y, Z] = \mathbf{Pr}[X|Z]\mathbf{Pr}[Y|Z]\mathbf{Pr}[Z].$$

- **Notation:** $X \perp\!\!\!\perp Y|Z$

Mean, variance, covariance

- For discrete RVs

$$\mathbb{E}[X] = \sum_x x \mathbf{Pr}[X = x]$$

and for continuous

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

- The variance and the standard deviation $\mathrm{std}[X] = \sigma$ are defined as

$$\mathbb{V}ar[X] = \sigma^2 = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2.$$

- Reminder: Jensen's inequality states that if f is convex, then

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)].$$

Mean, variance, covariance

Covariance of two random variables X, Y

$$\mathrm{cov}[X, Y] = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right] =$$
$$= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y].$$

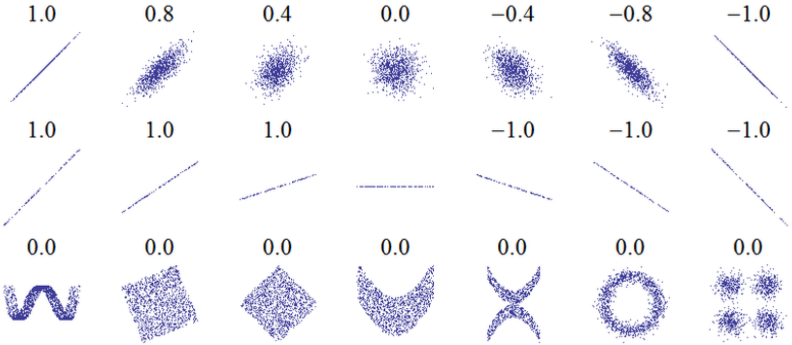
In general, if x is a d-dimensional random vector, the covariance is defined as

$$\mathrm{cov}[x] = \mathbb{E}\left[(x - \mathbb{E}[x])(x - \mathbb{E}[x])^T\right].$$

Pearson correlation coefficient:

$$\mathrm{corr}[X, Y] = \frac{\mathrm{cov}[X, Y]}{\sqrt{\mathbb{V}ar[X]\,\mathbb{V}ar[Y]}}.$$

Mean, variance, covariance



Correlation examples, Wikipedia

Probability distributions

Distribution	SciPy name
beta	beta
binomial	binom
Cauchy	cauchy
chi-squared	chi2
exponential	expon
F	f
gamma	gamma
geometric	geom
hypergeometric	hypergeom
inverse gamma	invgamma
log-normal	lognorm
logistic	logistic
negative binomial	nbinom
normal	norm
Poisson	poisson
Student t	t
uniform	unif
Weibull	exponweib

Source We will go over few important ones.

Discrete distributions

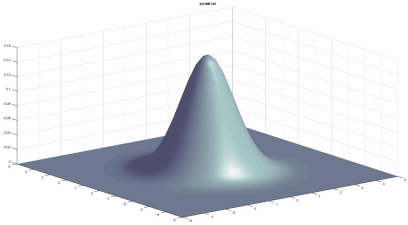
Details on whiteboard.

- **Bernoulli**: $X \sim Ber(p)$
- **Binomial**: $X \sim Bin(n, p)$
- **Multinomial**: $x \sim Mu(n, \theta)$
- **Poisson**: $X \sim Po(\lambda)$

Continuous Univariate distributions

- **Normal**: $X \sim N(x; \mu, \sigma^2)$
- **Student t distribution**: $X \sim \mathcal{T}(x; \mu, \sigma^2, \nu)$
- **Laplace**: $X \sim Lap(x; \mu, \beta)$
- **Gamma**: $X \sim Ga(x; \alpha, \beta)$
- **Pareto**: $\text{Pareto}(x|k, m)$

Multivariate normal distribution

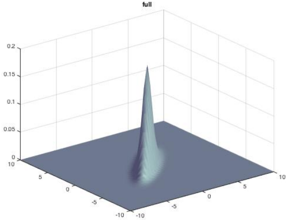


Isotropic, i.e., $\Sigma = \sigma^2 I$

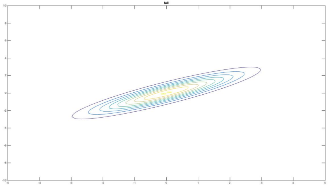
$$\mathcal{N}(x; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \right\}$$

where $\mu = \mathbb{E}[x]$, $\Sigma = \mathrm{Cov}[x]$. $\Sigma^{-1} = \Lambda$ is also known as the precision matrix.

Multivariate normal distribution



$$\mu = (0, 0), \Sigma = [21.8; 1.82]$$



Contour plot

Linear transformations of Random Variables

Suppose f is a linear function:

$$y = f(x) = Ax + b$$

Then,

$$\mathbb{E}[y] = A\mathbb{E}[x] + b \tag{6}$$

by Linearity of Expectation

$$\mathrm{Cov}[y] = A\mathrm{Cov}[x]A^T \tag{7}$$

Covariance

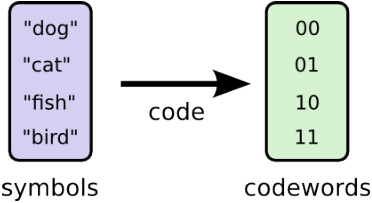
$$\mathrm{Cov}[y] = \mathbb{V}ar\left[a^T x + b\right] = a^T \mathrm{Cov}[x] a \tag{8}$$

if f() scalar valued

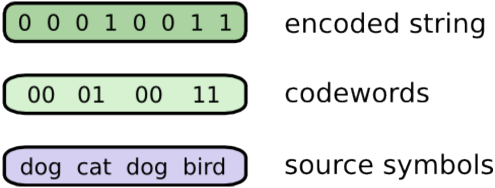
Information Theory

Information Theory

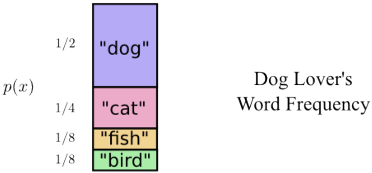
Suppose Bob wants to communicate with Alice by sending her bits.

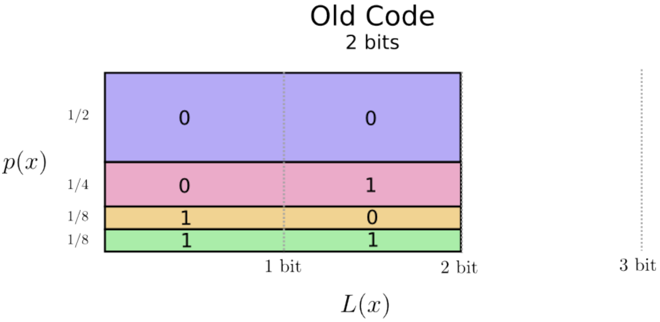


Example:



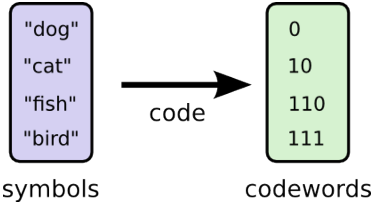
Information Theory

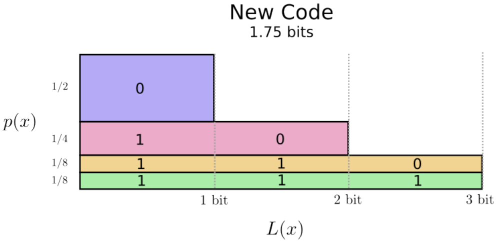




Can we use fewer than 2 bits?

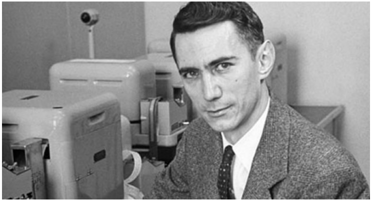
Information Theory





Can we use fewer than 1.75 bits?

Information Theory



- Suppose there are n events, the k-th event with probability p_k
- Shannon entropy, or just entropy is defined as:

$$\boxed{H(p_1, \ldots, p_n) = \sum_{k=1}^{n} p_k \log_2(\frac{1}{p_k}).}$$

Information Theory

Intuition:

- When the k-th event happends, we receive $\log(\frac{1}{p_k})$ bits of information.
- Therefore, $H(p_1, \ldots, p_n)$ is the expected number of bits in a random event.
- If $p_k = 0$, we define $p_k \log(\frac{1}{p_k}) = 0$

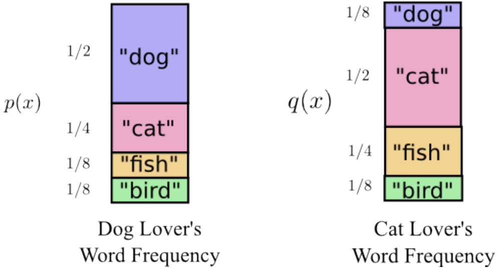
  To see why:

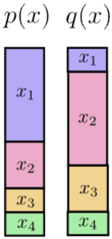
$$\lim_{\epsilon \to 0+} \epsilon \log(\frac{1}{p_k}) = 0.$$

- **Question:** For what values $p_1, \ldots, p_n$ is the entropy maximized?

Information Theory

Cross-entropy:





Cross-Entropy: $H_p(q)$

Average Length
of message from q(x)
using code for p(x).

Information Theory

Cross-entropy:

$$H_p(q) = \sum_x q(x) \log(\frac{1}{p(x)}).$$

- $H(p) = 1.75$

- $H(q) = 1.75$

- $H_p(q) = 2.25 \neq 2.375 = H_q(p)$

Cross-entropy isnt symmetric!

For the interested: Cross entropy and neural networks

Information Theory

Kullback-Leibler divergence (aka as relative entropy):

$$\mathrm{KL}(p, q) = \sum_k p_k \log(\frac{p_k}{q_k}).$$

$$\mathrm{KL}(p, q) = -H(p) + H_q(p).$$

Theorem

$$\mathrm{KL}(p, q) \geq 0 \tag{9}$$

Information Inequality

with equality iff $p = q$.

Information Theory

How similar is the joint probability distribution $p(X, Y)$ to the factorization $p(X)p(Y)$?

$$I(X; Y) = \mathrm{KL}(p(X, Y) || p(X)p(Y)) = \sum_{x,y} p(x, y) \log(\frac{p(x, y)}{p(x)p(y)}) \tag{10}$$

Mutual information