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

\[ p(A) = \frac{N_A}{N} \]
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.
- $X = \bigcup_i A_i$

Details on whiteboard.

$$|X|/52! = (52!)^{-1} \left( (\binom{52}{1}51! - (\binom{52}{2})50! + (\binom{52}{3})49! - \ldots - (\binom{52}{52})0! \right)$$

$$= 1 - 1/2! + 1/3! - \ldots - 1/52!$$

$$\approx 1 - \left( \sum_{i=0}^{\infty} (-1)^i / i! \right)$$

$$= 1 - 1/e.$$ 

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

Babis Tsourakakis
CS 591 Data Analytics, Lecture 2
Fundamental Rules

\[ \Pr [X \lor Y] = \Pr [X] + \Pr [Y] - \Pr [X \land Y] \]  \quad (1)

\[ \Pr [X] = \sum_y \Pr [X, Y = y] = \sum_y \Pr [X|Y = y] \Pr [Y = y] \]  \quad (2)

**Sum Rule**

\[ \Pr [X, Y] = \Pr [X \land Y] = \Pr [X] \Pr [Y|X] = \Pr [Y] \Pr [X|Y] \]  \quad (3)

**Product Rule**
Fundamental Rules

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

\[
\Pr[X_1, X_2, \ldots, X_n] = \Pr[X_1] \Pr[X_2, \ldots, X_n | X_1] = \ldots = \\
\Pr[X_1] \Pr[X_2 | X_1] \Pr[X_3 | X_2, X_1] \ldots \Pr[X_n | X_1, \ldots, X_{n-1}] \tag{4}
\]

Chain Rule

\[
\Pr[X | Y] = \frac{\Pr[X \land Y]}{\Pr[Y]} \tag{5}
\]

Conditional probability
Reminder: Bayes’ rule

Bayes’ rule is a direct application of conditional probabilities.

\[ \Pr[H|D] = \frac{\Pr[D|H]\Pr[H]}{\Pr[D]}, \text{ and } \Pr[D] > 0, \text{ or } \]

posterior \( \propto \) likelihood \( \times \) prior.
Independence and Conditional Independence

- We say $X$ and $Y$ are *unconditionally independent* or *marginally independent*, or just *independent* if
  \[
  \Pr[X \mid Y] = \Pr[X], \Pr[Y \mid X] = \Pr[Y]
  \]
- As a result
  \[
  \Pr[X, Y] = \Pr[X] \Pr[Y].
  \]
- **Notation:** $X \perp \perp Y$
Independence and Conditional Independence

Independence and Conditional Independence

Independence visualized

- **Weather next week**
  - Raining: 25%
  - Sunny: 75%

- **Clothing today**
  - T-shirt: 62%
  - Coat: 38%
Independence and Conditional Independence

Closer to reality

weather today

raining 25%
sunny 75%

clothing today
t-shirt 62%
coat 38%
Independence and Conditional Independence

Closer to reality ...

- Raining: 25%
- Sunny: 75%

- T-shirt: 25%
  - 6%
  - 56%

- Coat: 75%
  - 19%
  - 19%
Independence and Conditional Independence

... or alternatively ...

- Raining: 8% (6% probability given sunny)
- Sunny: 92% (56% probability given raining)
- Raining: 50% (19% probability given sunny)
- Sunny: 50% (19% probability given raining)

- T-shirt: 62%
- Coat: 38%
Independence and Conditional Independence

... or alternatively ...

raining

- 8%
- 6%
- 19%
- 56%
- sunny
- 92%
- t-shirt
- 62%
- coat
- 38%

raining & t-shirt
- 6%
- raining & coat
- 19%
- sunny & t-shirt
- 56%
- sunny & coat
- 19%
Independence and Conditional Independence

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

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

- Joint distribution factorizes as

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

- **Notation:** $X \perp \perp Y| Z$
Mean, variance, covariance

- For discrete RVs

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

and for continuous

\[ \mathbb{E} [X] = \int_x xp(x) \, dx \]

- The variance and the standard deviation \( \text{std}[X] = \sigma \) are defined as

\[ \text{Var} [X] = \sigma^2 = \mathbb{E} [(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2. \]

- **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 $X, Y$

$$\text{cov}[X, Y] = \mathbb{E} [(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] =$$

$$= \mathbb{E} [XY] - \mathbb{E}[X] \mathbb{E}[Y].$$

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

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

Pearson correlation coefficient:

$$\text{corr}[X, Y] = \frac{\text{cov}[X, Y]}{\sqrt{\text{Var}[X] \text{Var}[Y]}}.$$
Mean, variance, covariance

Correlation examples, Wikipedia
## Probability distributions

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

**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**: \( X \sim T(x; \mu, \sigma^2, \nu) \)

- **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., $\Sigma = \sigma^2 I$

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

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

\[ \mu = (0, 0), \Sigma = [21.8; 1.82] \]
Linear transformations of Random Variables

Suppose $f$ is a linear function:

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

Then,

$$\mathbb{E}[y] = A\mathbb{E}[x] + b$$  \hspace{1cm} (6)

by Linearity of Expectation

$$\text{Cov}[y] = ACov[x]A^T$$  \hspace{1cm} (7)

Covariance

$$\text{Cov}[y] = \text{Var}[a^Tx + b] = a^T\text{Cov}[x]a$$  \hspace{1cm} (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?

Babis Tsourakakis CS 591 Data Analytics, Lecture 2
Information Theory

Can we use fewer than 1.75 bits?

Babis Tsourakakis

CS 591 Data Analytics, Lecture 2
Information Theory

• Suppose there are $n$ events, the $k$-th event with probability $p_k$

• Shannon entropy, or just entropy is defined as:

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

- When the $k$-th event happens, we receive $\log\left(\frac{1}{p_k}\right)$ 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 $p_k \log\left(\frac{1}{p_k}\right) = 0$

To see why:

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

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

Cross-entropy:

\[ H_p(q) \]

Cross-Entropy: The average length of message from q(x) using code for p(x).
Information Theory

Cross-entropy:

\[ H_p(q) = \sum_x q(x) \log \left( \frac{1}{p(x)} \right). \]

- \( H(p) = 1.75 \)
- \( H(q) = 1.75 \)
- \( H_p(q) = 2.25 \neq 2.375 = H_q(p) \)

Cross-entropy isn't symmetric!

For the interested: Cross entropy and neural networks
Information Theory

**Kullback-Leibler divergence** *(aka as relative entropy)*:

\[
KL(p, q) = \sum_k p_k \log \left( \frac{p_k}{q_k} \right).
\]

\[
KL(p, q) = -H(p) + H_q(p).
\]

**Theorem**

\[
KL(p, q) \geq 0 \quad \text{(9)}
\]

*Information Inequality*

with equality iff \( p = q \).
How similar is the joint probability distribution $p(X, Y)$ to the factorization $p(X)p(Y)$?

\[
I(X; Y) = KL(p(X, Y) \| p(X)p(Y)) = \sum_{x, y} p(x, y) \log \left( \frac{p(x, y)}{p(x)p(y)} \right)
\]  

(10)

Mutual information