Statistical significance and decision trees
Aside – Knowledge

• **Anecdotes** are often useless except as proofs by contradiction.
  • E.g., “I saw Google used as a verb” does **not** mean that Google is **always** a verb, just that it is **not always** a noun.

• **Shallow statistics** are often not enough to be truly meaningful.
  • E.g., “My ASR system is 95% accurate on my test data. Yours is only 90% accurate, idiot.”
    • What if the test data was **biased** to favor my system?
    • What if we only used a **very small** amount of data?

• We need a **test** to see if our statistics actually **mean** something.
Differences due to sampling

• We previously saw KL divergence measure how different two distributions are from each other.

• But what if their difference is due to randomness in sampling?

• How can we tell that a distribution is really different from another?
Hypothesis testing

- Often, we assume a **null hypothesis**, $H_0$, which states that two distributions are **the same** (i.e., come from the same underlying model, population, or phenomenon).

- We **reject** the null hypothesis if there is **insufficient evidence** to support it.
  - This is often our goal – e.g., if my ASR system beats yours by 5%, I want to show that this difference is **not** a random accident.
  - As scientists, we have to be very **careful** to not reject $H_0$ too hastily.
    - How can we ensure our **diligence**?
Confidence

• We stated that we reject $H_0$ if there is insufficient evidence to support it.
  • How do we determine what is sufficient evidence?

• Significance level $\alpha$ ($0 \leq \alpha \leq 1$) is the maximum probability that two distributions are identical allowing us to disregard $H_0$.
  • In practice, $\alpha \leq 0.05$. Usually, it’s much lower.
  • Confidence level is $\gamma = 1 - \alpha$
  • E.g., a confidence level of 95% ($\alpha = 0.05$) implies that we expect that our decision is correct 95% of the time, regardless of the test data.
The *t*-test

• The *t*-test is a method to compute if distributions are significantly different from one another.

• It is based on the mean ($\bar{x}$) and variance ($\sigma$) of $N$ samples.

• It compares $\bar{x}$ and $\sigma$ to $H_0$ which states that the samples are drawn from a distribution with a mean $\mu$.
  • There are actually several types of $t$-tests for different situations...

• If $t = \frac{\bar{x} - \mu}{\sqrt{\sigma^2 / N}}$ (the “t-statistic”) is large enough, we can reject $H_0$.  

An example would be nice...
Example of the \textit{t}-test: tails

- Imagine that the average IQ of a UofT student is 158.
- We sample $N = 200$ UofT students from DCS and find that $\bar{x} = 169$ and $\sigma^2 = 2600$.
- Are DCS students significantly \textbf{smarter} than their peers?

- We use a \textbf{‘one-tailed’} test because we want to see if DCS students measure significantly \textbf{higher}.
  - If we just wanted to see if DCS were significantly \textbf{different}, we’d use a \textbf{two-tailed} test.
Example of the $t$-test: freedom

- Imagine that the average IQ of a UofT student is 158.
- We sample $N = 200$ UofT students from DCS and find that $\bar{x} = 169$ and $\sigma^2 = 2600$.
- Are DCS students significantly smarter than their peers?

- Degrees of freedom (d.f.): *n.pl.* In this $t$-test, this is the sum of the number of observations in each group, minus 2 (because there are two groups).

- In our example, we have $N_{DCS} = 200$ for DCS students, but $N_{UofT} \approx \infty$ for the other group, so $d.f. = \infty$.
  - (see Manning & Schütze for details)
Example of the $t$-test

• Imagine that the average IQ of a UofT student is 158.
• We sample $N = 200$ UofT students from DCS and find that $\bar{x} = 169$ and $\sigma^2 = 2600$.
• Are DCS students significantly smarter than their peers?

• So $t = \frac{\bar{x} - \mu}{\sqrt{\sigma^2/N}} = \frac{169 - 158}{\sqrt{2600/200}} \approx 3.05$

• In a $t$-test table, we look up the minimum value of $t$ necessary to reject $H_0$ at $\alpha = 0.005$ (we want to be quite confident) for a 1-tailed test...
Example of the $t$-test

• So $t = \frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{N}}} = \frac{169 - 158}{\sqrt{\frac{2600}{200}}} \approx 3.05$

• In a **$t$-test table**, we look up the minimum value of $t$ necessary to reject $H_0$ at $\alpha = 0.005$, and find 2.576.
  • Since $3.05 > 2.576$, we can reject $H_0$ at the 99.5% level of confidence ($\gamma = 1 - \alpha = 0.995$); **DCS students are significantly smarter**.
Example of the \( t \)-test

- Some things to observe about the \( t \)-test table:
  - We need **more evidence, \( t \)**, if we want to be **more confident** (left-right dimension).
  - We need **more evidence, \( t \)**, if we have **fewer measurements** (top-down dimension).
  - A common criticism of the \( t \)-test is that picking \( \alpha \) is ad-hoc. There are ways to correct for the selection of \( \alpha \).

<table>
<thead>
<tr>
<th>d.f.</th>
<th>( \alpha ) (one-tail)</th>
<th>0.05</th>
<th>0.025</th>
<th>0.01</th>
<th>0.005</th>
<th>0.001</th>
<th>0.0005</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.314</td>
<td>12.71</td>
<td>31.82</td>
<td>63.66</td>
<td>318.3</td>
<td>636.6</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1.812</td>
<td>2.228</td>
<td>2.764</td>
<td>3.169</td>
<td>4.144</td>
<td>4.587</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>1.725</td>
<td>2.086</td>
<td>2.528</td>
<td>2.845</td>
<td>3.552</td>
<td>3.850</td>
<td></td>
</tr>
<tr>
<td>( \infty )</td>
<td>1.645</td>
<td>1.960</td>
<td>2.326</td>
<td><strong>2.576</strong></td>
<td>3.091</td>
<td>3.291</td>
<td></td>
</tr>
</tbody>
</table>
Confidence of collocations

• **Collocation:** *n.* a ‘turn-of-phrase’ or usage where a sequence of words is *perceived* to have a meaning ‘*beyond*’ the sum of its parts.

• E.g., ‘*disk drive*’, ‘*video recorder*’, and ‘*soft drink*’ are collocations. ‘*cylinder drive*’, ‘*video measurer*’, ‘*weak drink*’ are *not* despite some near-synonymy between alternatives.

• Collocations are *not* just highly frequent bigrams, otherwise ‘*of the*’, and ‘*and the*’ would be collocations.

• How can we test if a bigram is a collocation or not?
Hypothesis testing collocations

• For collocations, our null hypothesis, $H_0$, is that there is no association between two given words beyond pure chance.
  • We compute the probability of those words occurring together if $H_0$ were true. If that probability is too low, we reject $H_0$.

• E.g., we expect ‘of the’ to occur together, because they’re both likely words to draw randomly
  • We could probably not reject $H_0$ in that case.
Example of the $t$-test on collocations

- Is ‘new companies’ a collocation?
- In our corpus of 14,307,668 word tokens, $new$ appears 15,828 times and $companies$ appears 4,675 times.
- Our null hypothesis, $H_0$ is that they are independent, i.e.,

$$H_0: P(new \text{ companies}) = P(new)P(companies)$$

$$= \frac{15828}{14307668} \times \frac{4675}{14307668}$$

$$\approx 3.615 \times 10^{-7}$$
Example of the $t$-test on collocations

• The Manning & Schütze text claims that if the process of randomly generating bigrams follows a **Bernoulli distribution**.

  • i.e., assigning 1 whenever *new companies* appears and 0 otherwise gives $\mu = p = P(\text{new companies})$

  • For Bernoulli distributions, $\sigma^2 = p(1 - p)$. Manning & Schütze claim that we can assume $\sigma^2 = p(1 - p) \approx p$, since for most bigrams, $p$ is very small.
Example of the $t$-test

- So, $\mu = 3.615 \times 10^{-7}$ is the expected mean in $H_0$.
- We **actually count** 8 occurrences of *new companies* in our corpus
  - $\bar{x} = \frac{8}{14307667} \approx 5.591 \times 10^{-7}$

- So $t = \frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{N}}} = \frac{5.591 \times 10^{-7} - 3.615 \times 10^{-7}}{\sqrt{\frac{5.591 \times 10^{-7}}{14307667}}} \approx 0.9999$
  - $\therefore \sigma^2 \approx p = \bar{x} = 5.591 \times 10^{-7}$

- In a **$t$-test table**, we look up the minimum value of $t$ necessary to reject $H_0$ at $\alpha = 0.005$, and find **2.576**.
  - Since **0.9999** < **2.576**, we cannot reject $H_0$ at the 99.5% level of confidence.
    - We **don’t have enough evidence** to think that *new companies* is a collocation (we can’t say that it definitely *isn’t*, though!).
Aside – analysis of variance

- **Analyses of variance (ANOVAs)** (there are several types) can be:
  - A way to **generalize t-tests** to more than two groups.
  - A way to **determine which** (if any) of several **variables** are responsible for the **variation** in an observation (and the interaction between them).

- E.g., we measure the **accuracy** of an ASR system for different settings of **empirical parameters** $M$ and $Q$ (more on these later in the course...).

<table>
<thead>
<tr>
<th>Accuracy (%)</th>
<th>$M = 2$</th>
<th>$M = 4$</th>
<th>$M = 16$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Q = 2$</td>
<td>53.33</td>
<td>66.67</td>
<td>53.33</td>
</tr>
<tr>
<td></td>
<td>26.67</td>
<td>53.33</td>
<td>40.00</td>
</tr>
<tr>
<td>$Q = 5$</td>
<td>93.33</td>
<td>26.67</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>66.67</td>
<td>13.33</td>
<td>80.00</td>
</tr>
<tr>
<td></td>
<td>40.00</td>
<td>0.00</td>
<td>60.00</td>
</tr>
</tbody>
</table>

- **$H_0$: no effect of source variables.**

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>$p$ value</th>
<th>Accept/Reject $H_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Q$</td>
<td>1</td>
<td>0.179</td>
<td>Accept $H_0$</td>
</tr>
<tr>
<td>$M$</td>
<td>2</td>
<td>0.106</td>
<td>Accept $H_0$</td>
</tr>
<tr>
<td>interaction</td>
<td>2</td>
<td>0.006</td>
<td>Reject $H_0$ at $\alpha = 0.01$</td>
</tr>
</tbody>
</table>

A completely fictional example
Trees!

(The larch.)
Decision trees

• Consists of rules for classifying data that consists of many attributes.

  • **Decision nodes:** Non-terminal. Consists of a question asked of one of the attributes, and a branch for each possible answer.

  • **Leaf nodes:** Terminal. Consists of a single class/category, so no further testing is required.
Decision tree example

• Shall I go for a walk?

```
Forecast
  - SUNNY
    - Humidity
      - HIGH
        - NO!
      - LOW
        - YES!
  - RAIN
    - Windy
      - TRUE
        - NO!
      - FALSE
        - YES!
```
Decision tree algorithm: ID3

- **ID3** (iterative dichotomiser 3) is an algorithm invented by Ross Quinlan to produce decision trees from data.

- Basically,
  1. Compute the entropy of asking about each attribute.
  2. Choose the attribute which reduces the most entropy.
  3. Make a node asking a question of that attribute.
  4. Go to step 1, minus the chosen attribute.

- Example attribute vectors (observations):

<table>
<thead>
<tr>
<th>Forecast</th>
<th>Humidity</th>
<th>Wind</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. token length</td>
<td>Avg. sentence length</td>
<td>Frequency of nouns</td>
</tr>
</tbody>
</table>
Information gain

• The **information gain** is based on the expected decrease in entropy after a set of **training** data is split on an attribute.
• We prefer the attribute that removes the most entropy.

\[
Gain(Q) = H(S) - \sum_{\text{child set}} p(\text{child set})H(\text{child set})
\]

\[
S = A \cup B \\
\emptyset = A \cap B
\]

Each of \( S \), \( A \), and \( B \) consist of examples from the data
Information gain and ID3

• When a node in the decision tree is \textit{generated} in which \textbf{all members} have a \textbf{common} class,
  • that node has 0 entropy,
  • that node is a leaf node.

• Otherwise, we need to (try to) split that node with another question.
### Example – Gender classification

<table>
<thead>
<tr>
<th>Person</th>
<th>Hair Length</th>
<th>Mass</th>
<th>Age</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homer</td>
<td>0”</td>
<td>250</td>
<td>36</td>
<td>M</td>
</tr>
<tr>
<td>Marge</td>
<td>10”</td>
<td>150</td>
<td>34</td>
<td>F</td>
</tr>
<tr>
<td>Bart</td>
<td>2”</td>
<td>90</td>
<td>10</td>
<td>M</td>
</tr>
<tr>
<td>Lisa</td>
<td>6”</td>
<td>78</td>
<td>8</td>
<td>F</td>
</tr>
<tr>
<td>Maggie</td>
<td>4”</td>
<td>20</td>
<td>1</td>
<td>F</td>
</tr>
<tr>
<td>Abe</td>
<td>1”</td>
<td>170</td>
<td>70</td>
<td>M</td>
</tr>
<tr>
<td>Selma</td>
<td>8”</td>
<td>160</td>
<td>41</td>
<td>F</td>
</tr>
<tr>
<td>Otto</td>
<td>10”</td>
<td>180</td>
<td>38</td>
<td>M</td>
</tr>
<tr>
<td>Krusty</td>
<td>6”</td>
<td>200</td>
<td>45</td>
<td>M</td>
</tr>
<tr>
<td>Comic</td>
<td>8”</td>
<td>290</td>
<td>38</td>
<td>?</td>
</tr>
</tbody>
</table>
Split on hair length?

- **YES** Hair Length ≤ 5?
- **NO**

\[
H(S) = \frac{m}{m+f} \log_2 \left( \frac{m+f}{m} \right) + \frac{f}{m+f} \log_2 \left( \frac{m+f}{f} \right)
\]

\[
H(4F, 5M) = \frac{4}{9} \log_2 \left( \frac{9}{4} \right) + \frac{5}{9} \log_2 \left( \frac{9}{5} \right) = 0.9911 \text{ bits}
\]

\[
H(3F, 2M) = \frac{3}{5} \log_2 \left( \frac{5}{3} \right) + \frac{2}{5} \log_2 \left( \frac{5}{2} \right) = 0.971
\]

\[
H(1F, 3M) = \frac{1}{4} \log_2 \left( \frac{4}{1} \right) + \frac{3}{4} \log_2 \left( \frac{4}{3} \right) = 0.8113
\]

- **Gain(Question)** = \( H(S) - \sum_{child\ set} p(\text{child set}) H(\text{child set}) \)
- **Gain(HairLen ≤ 5)** = 0.9911 − \( \frac{4}{9} \times 0.8113 \) − \( \frac{5}{9} \times 0.971 = 0.0911 \)
Split on mass?

$H(S) = \frac{m}{m+f} \log_2 \left( \frac{m+f}{m} \right) + \frac{f}{m+f} \log_2 \left( \frac{m+f}{f} \right)$

$H(4F, 5M) = \frac{4}{9} \log_2 \left( \frac{9}{4} \right) + \frac{5}{9} \log_2 \left( \frac{9}{5} \right) = 0.9911 \text{ bits}$

$H(0F, 4M) = \frac{0}{4} \log_2(\infty) + \frac{4}{4} \log_2 \left( \frac{4}{4} \right) = 0$

$H(4F, 1M) = \frac{4}{5} \log_2 \left( \frac{5}{4} \right) + \frac{1}{5} \log_2 \left( \frac{5}{1} \right) = 0.7219$

- $Gain(\text{Question}) = H(S) - \sum_{\text{child set}} p(\text{child set}) H(\text{child set})$
- $Gain(\text{Mass} \leq 160) = 0.9911 - \frac{5}{9} \cdot 0.7219 - \frac{4}{9} \cdot 0 = 0.59$
Split on age?

\[ H(S) = \frac{m}{m+f} \log_2 \left( \frac{m+f}{m} \right) + \frac{f}{m+f} \log_2 \left( \frac{m+f}{f} \right) \]

\[ H(4F, 5M) = \frac{4}{9} \log_2 \left( \frac{9}{4} \right) + \frac{5}{9} \log_2 \left( \frac{9}{5} \right) = 0.9911 \text{ bits} \]

\[ H(1F, 2M) = \frac{1}{3} \log_2 \left( \frac{3}{1} \right) + \frac{2}{3} \log_2 \left( \frac{3}{2} \right) = 0.9183 \]

\[ H(3F, 3M) = \frac{3}{6} \log_2 \left( \frac{6}{3} \right) + \frac{3}{6} \log_2 \left( \frac{6}{3} \right) = 1 \]

• Gain(Question) = \( H(S) - \sum_{\text{child set}} p(\text{child set}) H(\text{child set}) \)
• Gain(Age \leq 40) = 0.9911 - \frac{6}{9} \cdot 1 - \frac{3}{9} \cdot 0.9183 = 0.0183
The resulting tree

- Splitting on mass resulted in the greatest information gain.
- We’re left with one heterogeneous set, so we recurse and find that hair length results in a complete classification of the training data.
Testing

- We just need to keep track of the attribute questions – not the training data.

- How are the following characters classified?

<table>
<thead>
<tr>
<th>Person</th>
<th>Hair length</th>
<th>Mass</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comic</td>
<td>8”</td>
<td>290</td>
<td>38</td>
</tr>
<tr>
<td>Hans</td>
<td>0”</td>
<td>100</td>
<td>80</td>
</tr>
<tr>
<td>Janey</td>
<td>7”</td>
<td>74</td>
<td>8</td>
</tr>
<tr>
<td>Jimbo</td>
<td>6”</td>
<td>140</td>
<td>16</td>
</tr>
</tbody>
</table>

- Thanks to Allan Neymark (San Jose State University) for Simpsons example.
Aspects of ID3

- ID3 tends to build **short trees** since at each step we are removing the maximum amount of entropy possible.
- ID3 trains on the **whole training set** and does not succumb to issues related to **random initialization**.
- ID3 can **over-fit** to training data.
- Only **one attribute is used at a time** to make decisions.
- It can be difficult to use **continuous** data, since many trees need to be generated to see where to break the continuum.
Aspects of C4.5

• An extension to ID3 by the same creator, Ross Quinlan.
• Can support continuous attributes.
• Can support non-binary decision.
• Can **prune** the tree after creation to simplify it.
• Supports cross-validation.
• Is merely a ‘black-box’ for this course (called J48).
Next week

• Hidden Markov models!

Image sort of from 2001: A Space Odyssey by MGM pictures