Neural Net Demo!

https://playground.tensorflow.org/
Neural Networks
Multi-class Logistic Regression

- special case of neural network

\[
P(y_1|x; w) = \frac{e^{z_2}}{e^{z_1} + e^{z_2} + e^{z_3}}
\]

\[
P(y_2|x; w) = \frac{e^{z_2}}{e^{z_1} + e^{z_2} + e^{z_3}}
\]

\[
P(y_3|x; w) = \frac{e^{z_3}}{e^{z_1} + e^{z_2} + e^{z_3}}
\]
Deep Neural Network = Also learn the features!

\[ f_1(x) \]
\[ f_2(x) \]
\[ f_3(x) \]
\[ \ldots \]
\[ f_n(x) \]
\[ z_1 \]
\[ z_2 \]
\[ z_3 \]
\[ \text{softmax} \]
\[ P(y_1|x; w) = \frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}} \]
\[ P(y_2|x; w) = \frac{e^{z_2}}{e^{z_1} + e^{z_2} + e^{z_3}} \]
\[ P(y_3|x; w) = \frac{e^{z_3}}{e^{z_1} + e^{z_2} + e^{z_3}} \]
Deep Neural Network = Also learn the features!

\[
z_i^{(k)} = g\left( \sum_j W_{i,j}^{(k-1,k)} z_j^{(k-1)} \right)
\]

\(g = \text{nonlinear activation function}\)
Deep Neural Network = Also learn the features!

\[ z_i^{(k)} = g\left( \sum_j W_{i,j}^{(k-1,k)} z_j^{(k-1)} \right) \]

\[ g = \text{nonlinear activation function} \]
Common Activation Functions

Sigmoid Function

\[ g(z) = \frac{1}{1 + e^{-z}} \]
\[ g'(z) = g(z)(1 - g(z)) \]

Hyperbolic Tangent

\[ g(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \]
\[ g'(z) = 1 - g(z)^2 \]

Rectified Linear Unit (ReLU)

\[ g(z) = \max(0, z) \]
\[ g'(z) = \begin{cases} 1, & z > 0 \\ 0, & \text{otherwise} \end{cases} \]
Training the deep neural network is just like logistic regression:

\[
\max_w \ ll(w) = \max_w \sum_i \log P(y^{(i)} | x^{(i)}; w)
\]

just \(w\) tends to be a much, much larger vector 😊

→ just run gradient ascent

+ stop when log likelihood of hold-out data starts to decrease
Neural Networks Properties

- **Theorem (Universal Function Approximators).** A two-layer neural network with a sufficient number of neurons can approximate any continuous function to any desired accuracy.

- **Practical considerations**
  - Can be seen as learning the features
  - Large number of neurons
    - Danger for overfitting
    - (hence early stopping!)
How about computing all the derivatives?

- Derivatives tables:

\[
\frac{d}{dx}(a) = 0 \\
\frac{d}{dx}(x) = 1 \\
\frac{d}{dx}(au) = a \frac{du}{dx} \\
\frac{d}{dx}(u + v - w) = \frac{du}{dx} + \frac{dv}{dx} - \frac{dw}{dx} \\
\frac{d}{dx}(uv) = u \frac{dv}{dx} + v \frac{du}{dx} \\
\frac{d}{dx} \left( \frac{u}{v} \right) = \frac{1}{v} \frac{du}{dx} - \frac{u}{v^2} \frac{dv}{dx} \\
\frac{d}{dx}(u^n) = nu^{n-1} \frac{du}{dx} \\
\frac{d}{dx}(\sqrt{u}) = \frac{1}{2\sqrt{u}} \frac{du}{dx} \\
\frac{d}{dx} \left( \frac{1}{u} \right) = -\frac{1}{u^2} \frac{du}{dx} \\
\frac{d}{dx} \left( \frac{1}{u^n} \right) = -\frac{n}{u^{n+1}} \frac{du}{dx} \\
\frac{d}{dx}[f(u)] = \frac{df}{du} \frac{du}{dx} \\
\frac{d}{dx}[\ln(u)] = \frac{d}{dx}[\log_a u] = \frac{1}{u} \frac{du}{dx} \\
\frac{d}{dx}[\log_a u] = \frac{d}{dx}[\log_e u] = \frac{1}{u} \frac{du}{dx} \\
\frac{d}{dx}[a^u] = a^u \ln a \frac{du}{dx} \\
\frac{d}{dx}[e^u] = e^u \frac{du}{dx} \\
\frac{d}{dx}[\sin u] = \cos u \frac{du}{dx} \\
\frac{d}{dx}[\cos u] = -\sin u \frac{du}{dx} \\
\frac{d}{dx}[\tan u] = \sec^2 u \frac{du}{dx} \\
\frac{d}{dx}[\cot u] = -\csc^2 u \frac{du}{dx} \\
\frac{d}{dx}[\sec u] = \sec u \tan u \frac{du}{dx} \\
\frac{d}{dx}[\csc u] = -\csc u \cot u \frac{du}{dx}
\]

[source: http://hyperphysics.phy-astr.gsu.edu/hbase/Math/derfunc.html]
How about computing all the derivatives?

But neural net $f$ is never one of those?

No problem: CHAIN RULE:

If \[ f(x) = g(h(x)) \]

Then \[ f'(x) = g'(h(x))h'(x) \]

→ Derivatives can be computed by following well-defined procedures.
Automatic Differentiation

- **Automatic differentiation software**
  - e.g. Theano, TensorFlow, PyTorch, Chainer
  - Only need to program the function $g(x,y,w)$
  - Can automatically compute all derivatives w.r.t. all entries in $w$
  - This is typically done by caching info during forward computation pass of $f$, and then doing a backward pass = “backpropagation”
  - Autodiff / Backpropagation can often be done at computational cost comparable to the forward pass

- Need to know this exists
- How this is done? -- outside of scope of CS188
Training a Network (setting weights)

\[ z_i^{(k)} = g \left( \sum_j W_{i,j}^{(k-1,k)} z_j^{(k-1)} \right) \]

\[ z = \text{nonlinear activation function} \]
Training a Network

Key words:
- Forward
- Backwards
- Gradient
- Backprop

\[ P(y_1|x;w) = e^{z_{1(OUT)}} \]
\[ P(y_2|x;w) = e^{z_{2(OUT)}} \]
\[ P(y_3|x;w) = e^{z_{3(OUT)}} \]

\[ g = \text{nonlinear activation function} \]
Suppose we have $g(w) = w_1^3w_2 + 3w_1$ and want the gradient at $w = [2, 3]$.

Think of the function as a composition of many functions.

- Can use derivative chain rule to compute $\frac{\partial g}{\partial w_1}$ and $\frac{\partial g}{\partial w_2}$.

$g = b + c$

- $\frac{\partial g}{\partial b} = 1$, $\frac{\partial g}{\partial c} = 1$
Back Propagation: \( g(\mathbf{w}) = w_1^3 w_2 + 3w_1 \)

- Suppose we have \( g(\mathbf{w}) = w_1^3 w_2 + 3w_1 \) and want the gradient at \( \mathbf{w} = [2, 3] \).
- Think of the function as a composition of many functions.
  - Can use derivative chain rule to compute \( \frac{\partial g}{\partial w_1} \) and \( \frac{\partial g}{\partial w_2} \).
- \( g = b + c \)
  - \( \frac{\partial g}{\partial b} = 1, \frac{\partial g}{\partial c} = 1 \)
- \( b = a \times w_2 \)
  - \( \frac{\partial g}{\partial a} = \frac{\partial g}{\partial b} \times \frac{\partial b}{\partial a} \)

\[
\begin{align*}
g &= b + c \\
b &= a \times w_2
\end{align*}
\]
Suppose we have \( g(\mathbf{w}) = w_1^3w_2 + 3w_1 \) and want the gradient at \( \mathbf{w} = [2, 3] \).

Think of the function as a composition of many functions.

- Can use derivative chain rule to compute \( \frac{\partial g}{\partial w_1} \) and \( \frac{\partial g}{\partial w_2} \).

\[
\begin{align*}
g &= b + c \\
\frac{\partial g}{\partial b} &= 1, \quad \frac{\partial g}{\partial c} = 1 \\
b &= a \times w_2 \\
\frac{\partial g}{\partial a} &= \frac{\partial g}{\partial b} \frac{\partial b}{\partial a} = \frac{\partial g}{\partial w_1} \frac{\partial w_1}{\partial a} = \frac{\partial g}{\partial w_1} \end{align*}
\]
Suppose we have \( g(w) = w_1^3 w_2 + 3 w_1 \) and want the gradient at \( w = [2, 3] \).

Think of the function as a composition of many functions.

- Can use derivative chain rule to compute \( \frac{\partial g}{\partial w_1} \) and \( \frac{\partial g}{\partial w_2} \).

- \( g = b + c \)
  - \( \frac{\partial g}{\partial b} = 1, \frac{\partial g}{\partial c} = 1 \)

- \( b = a \times w_2 \)
  - \( \frac{\partial g}{\partial a} = \frac{\partial g}{\partial b} \frac{\partial b}{\partial a} = 1 \frac{\partial b}{\partial a} = 1 \cdot 3 = 3 \)

\[ g = b + c = a + b + c = 30 \]

\[ \frac{\partial g}{\partial a} = \frac{\partial g}{\partial b} = 1 \quad \frac{\partial g}{\partial c} = 1 \]

\[ \frac{\partial g}{\partial w_1} = 3 + 3 = 6 \quad \frac{\partial g}{\partial w_2} = 3 \]

\[ \frac{\partial g}{\partial w} = [6, 3] \]
Back Propagation: \( g(\mathbf{w}) = w_1^3 w_2 + 3w_1 \)

- Suppose we have \( g(\mathbf{w}) = w_1^3 w_2 + 3w_1 \) and want the gradient at \( \mathbf{w} = [2, 3] \).
- Think of the function as a composition of many functions.
  - Can use derivative chain rule to compute \( \frac{\partial g}{\partial w_1} \) and \( \frac{\partial g}{\partial w_2} \).

\[
\begin{align*}
  g &= b + c \\
  \frac{\partial g}{\partial b} &= 1, \quad \frac{\partial g}{\partial c} = 1 \\
  b &= a \times w_2 \\
  \frac{\partial g}{\partial a} &= \frac{\partial g}{\partial b} \frac{\partial b}{\partial a} = 1 \frac{db}{da} = 1 \cdot 3 = 3 \\
  a &= w_1^3 \\
  \frac{\partial g}{\partial w_1} &= \ldots
\end{align*}
\]
Suppose we have $g(w) = w_1^3 w_2 + 3w_1$ and want the gradient at $w = [2, 3]$. Think of the function as a composition of many functions.

- Can use derivative chain rule to compute $\frac{\partial g}{\partial w_1}$ and $\frac{\partial g}{\partial w_2}$.

$g = b + c$
- $\frac{\partial g}{\partial b} = 1, \frac{\partial g}{\partial c} = 1$

$b = a \times w_2$
- $\frac{\partial g}{\partial a} = \frac{\partial g}{\partial b} \frac{\partial b}{\partial a} = 1 \frac{\partial b}{\partial a} = 1 \cdot 3 = 3$

$a = w_1^3$
- $\frac{\partial g}{\partial w_1} = \frac{\partial g}{\partial a} \frac{\partial a}{\partial w_1} = 3 \cdot 3w_1^2 = 36$

Interpretation: A tiny increase in $w_1$ will result in an approximately $36w_1$ increase in $g$ due to this cube function.
Suppose we have \( g(\mathbf{w}) = w_1^3 w_2 + 3w_1 \) and want the gradient at \( \mathbf{w} = [2, 3] \).

Think of the function as a composition of many functions.

- Can use derivative chain rule to compute \( \frac{\partial g}{\partial w_1} \) and \( \frac{\partial g}{\partial w_2} \).

\[
\begin{align*}
g &= b + c \\
\frac{\partial g}{\partial b} &= 1, \quad \frac{\partial g}{\partial c} = 1 \\
b &= a \times w_2 \\
\frac{\partial g}{\partial a} &= \frac{\partial g}{\partial b} \frac{\partial b}{\partial a} = 1 \frac{\partial b}{\partial a} = 1 \cdot 3 = 3 \\
a &= w_1^3 \\
\frac{\partial g}{\partial w_1} &= \frac{\partial g}{\partial a} \frac{\partial a}{\partial w_1} = 3 \cdot 3w_1^2 = 36 \\
\frac{\partial g}{\partial w_2} &= ? \quad \text{Hint: } b = a \times 3 \text{ may be useful.}
\end{align*}
\]
Suppose we have \( g(w) = w_1^3 w_2 + 3w_1 \) and want the gradient at \( w = [2, 3] \).

Think of the function as a composition of many functions.

- Can use derivative chain rule to compute \( \frac{\partial g}{\partial w_1} \) and \( \frac{\partial g}{\partial w_2} \).

\( g = b + c \)

\[ \frac{\partial g}{\partial b} = 1, \quad \frac{\partial g}{\partial c} = 1 \]

\( b = a \times w_2 \)

\[ \frac{\partial g}{\partial a} = \frac{\partial g}{\partial b} \frac{\partial b}{\partial a} = 1 \cdot 3 = 3 \]

\[ \frac{\partial g}{\partial w_2} = \frac{\partial g}{\partial b} \frac{\partial b}{\partial w_2} = \frac{1 \cdot 8}{1} = 8 \]

\( a = w_1^3 \)

\[ \frac{\partial g}{\partial w_1} = \frac{\partial g}{\partial a} \frac{\partial a}{\partial w_1} = 3 \cdot 3w_1^2 = 36 \]
Back Propagation: \( g(w) = w_1^3w_2 + 3w_1 \)

- Suppose we have \( g(w) = w_1^3w_2 + 3w_1 \) and want the gradient at \( w = [2, 3] \)
- Think of the function as a composition of many functions, use chain rule.
- \( g = b + c \)
  - \( \frac{\partial g}{\partial b} = 1, \frac{\partial g}{\partial c} = 1 \)
- \( b = a \times w_2 \)
  - \( \frac{\partial g}{\partial a} = \frac{\partial g}{\partial b} \frac{\partial b}{\partial a} = 1 \frac{\partial b}{\partial a} = 1 \cdot 3 = 3 \)
  - \( \frac{\partial g}{\partial w_2} = \frac{\partial g}{\partial b} \frac{\partial b}{\partial w_2} = 1 \frac{\partial b}{\partial w_2} = 1 \cdot 8 = 8 \)
- \( a = w_1^3 \)
  - \( \frac{\partial g}{\partial w_1} = \frac{\partial g}{\partial a} \frac{\partial a}{\partial w_1} = 3 \cdot 3w_1^2 = 36 \)
- \( c = 3w_1 \)
  - \( \frac{\partial g}{\partial w_1} = \frac{\partial g}{\partial c} \frac{\partial c}{\partial w_1} = 1 \cdot 3 = 3 \)

How do we reconcile this seeming contradiction?
Top partial derivative means cube function contributes 36\(w_1\) and bottom p.d. means product contributes 3\(w_1\) so add them.
Suppose we have $g(w) = w_1^3w_2 + 3w_1$ and want the gradient at $w = [2, 3]$. Think of the function as a composition of many functions, use chain rule.

- $g = b + c$
  - $\frac{dg}{db} = 1, \frac{dg}{dc} = 1$

- $b = a \times w_2$
  - $\frac{dg}{da} = \frac{dg}{db} \frac{db}{da} = 1 \frac{db}{da} = 1 \cdot 3 = 3$
  - $\frac{dg}{dw_2} = \frac{dg}{db} \frac{db}{dw_2} = 1 \frac{db}{dw_2} = 1 \cdot 8 = 8$

- $a = w_1^3$
  - $\frac{dg}{dw_1} = \frac{dg}{da} \frac{da}{dw_1} = 3 \cdot 3w_1^2 = 36$

- $c = 3w_1$
  - $\frac{dg}{dw_1} = \frac{dg}{dc} \frac{dc}{dw_1} = 1 \cdot 3 = 3$

The gradient at $w = [2, 3]$ is $\nabla g = \left[ \frac{dg}{dw_1}, \frac{dg}{dw_2} \right] = [39, 8]$. 

Diagram: 

- $w_1 = 2$ 
  - $w_1 = 3$ 
    - $w_1 = 36$ 
      - $w_1 = a = 8$ 
        - $w_1 = b = 24$ 
          - $w_1 = g = 30$ 
    - $w_1 = 8$ 
      - $w_1 = 3$ 
        - $w_1 = c = 6$ 
          - $w_1 = 1$ 
  - $w_1 = 3$ 
    - $w_1 = 2$ 
      - $w_1 = 3$ 
        - $w_1 = 6$ 
          - $w_1 = 1$
Gradient Descent

- **Punchline:** If we can somehow compute our gradient, we can use gradient descent.
- **How do we compute the gradient?**
  - **Purely analytically.**
    - Gives exact symbolic answer. Infeasible for functions of lots of parameters or input values.
  - **Finite difference approximation.**
    - Gives approximation, very easy to implement.
    - Runtime for \( \|
  \ll \|
 \): \( O(NM) \), where \( N \) is the number of parameters, and \( M \) is number of data points.
  - **Back propagation.**
    - Gives exact answer, difficult to implement.
    - Runtime for \( \|
  \ll \|
 \): \( O(NM) \)

\[
ll(w) = \sum_{i=1}^{m} \log p(y = y^{(i)} | f(x^{(i)}) ; w)
\]
Optimize probability of label given input

\[ \max_w ll(w) = \max_w \sum_i \log P(y^{(i)} | x^{(i)}; w) \]

Continuous optimization

- Gradient ascent:
  - Compute steepest uphill direction = gradient (= just vector of partial derivatives)
  - Take step in the gradient direction
  - Repeat (until held-out data accuracy starts to drop = “early stopping”)

Deep neural nets

- Last layer = still logistic regression
- Now also many more layers before this last layer
  - = computing the features
  - \( \rightarrow \) the features are learned rather than hand-designed

Universal function approximation theorem

- If neural net is large enough
- Then neural net can represent any continuous mapping from input to output with arbitrary accuracy
- But remember: need to avoid overfitting / memorizing the training data \( \rightarrow \) early stopping!

Automatic differentiation gives the derivatives efficiently (how? = outside of scope of 188)
Computer Vision
Manual Feature Design
Features and Generalization

Image

HoG
Speech Recognition

TIMIT Speech Recognition

- Traditional
- Deep Learning

Error Rate


graph credit Matt Zeiler, Clarifai
What’s still missing? – correlation \neq causation

Figure 11: Raw data and explanation of a bad model’s prediction in the “Husky vs Wolf” task.

<table>
<thead>
<tr>
<th></th>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trusted the bad model</td>
<td>10 out of 27</td>
<td>3 out of 27</td>
</tr>
<tr>
<td>Snow as a potential feature</td>
<td>12 out of 27</td>
<td>25 out of 27</td>
</tr>
</tbody>
</table>

Table 2: “Husky vs Wolf” experiment results.

[Ribeiro et al.]
What’s still missing? – covariate shift

Orders: onion
Time Left: 297

[Carroll et al.]
What’s still missing? – covariate shift

[Carroll et al.]
Decision Trees
Reminder: Features

- Features, aka attributes
  - Sometimes: TYPE=French
  - Sometimes: $f_{TYPE=French}(x) = 1$

<table>
<thead>
<tr>
<th>Example</th>
<th>Alt</th>
<th>Bar</th>
<th>Fri</th>
<th>Hun</th>
<th>Pat</th>
<th>Price</th>
<th>Rain</th>
<th>Res</th>
<th>Type</th>
<th>Est</th>
<th>Target</th>
<th>WillWait</th>
</tr>
</thead>
<tbody>
<tr>
<td>X₁</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>Some</td>
<td>$$$</td>
<td>F</td>
<td>T</td>
<td>French</td>
<td>0–10</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>X₂</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>Full</td>
<td>$</td>
<td>F</td>
<td>F</td>
<td>Thai</td>
<td>30–60</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>X₃</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>Some</td>
<td>$</td>
<td>F</td>
<td>F</td>
<td>Burger</td>
<td>0–10</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>X₄</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>Full</td>
<td>$</td>
<td>F</td>
<td>F</td>
<td>Thai</td>
<td>10–30</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>X₅</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>Full</td>
<td>$$$</td>
<td>F</td>
<td>T</td>
<td>French</td>
<td>&gt;60</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>X₆</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>Some</td>
<td>$</td>
<td>T</td>
<td>T</td>
<td>Italian</td>
<td>0–10</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>X₇</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>None</td>
<td>$</td>
<td>T</td>
<td>F</td>
<td>Burger</td>
<td>0–10</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>X₈</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>Some</td>
<td>$</td>
<td>T</td>
<td>T</td>
<td>Thai</td>
<td>0–10</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>X₉</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>Full</td>
<td>$</td>
<td>T</td>
<td>F</td>
<td>Burger</td>
<td>&gt;60</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>X₁₀</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>Full</td>
<td>$$$</td>
<td>F</td>
<td>T</td>
<td>Italian</td>
<td>10–30</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>X₁₁</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>None</td>
<td>$</td>
<td>F</td>
<td>F</td>
<td>Thai</td>
<td>0–10</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>X₁₂</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>Full</td>
<td>$</td>
<td>F</td>
<td>F</td>
<td>Burger</td>
<td>30–60</td>
<td>T</td>
<td></td>
</tr>
</tbody>
</table>
Decision Trees

- Compact representation of a function:
  - Truth table
  - Conditional probability table
  - Regression values

- True function
  - Realizable: in $H$
Expressiveness of DTs

- Can express any function of the features

$P(C|A, B)$

- However, we hope for compact trees
Comparison: Perceptrons

- What is the expressiveness of a perceptron over these features?

- For a perceptron, a feature’s contribution is either positive or negative
  - If you want one feature’s effect to depend on another, you have to add a new conjunction feature
  - E.g. adding “PATRONS=full ∧ WAIT = 60” allows a perceptron to model the interaction between the two atomic features

- DTs automatically conjoin features / attributes
  - Features can have different effects in different branches of the tree!

- Difference between modeling relative evidence weighting (NB) and complex evidence interaction (DTs)
  - Though if the interactions are too complex, may not find the DT greedily
Decision Tree Learning

- **Aim:** find a small tree consistent with the training examples
- **Idea:** (recursively) choose “most significant” attribute as root of (sub)tree

```plaintext
function DTL(examples, attributes, default) returns a decision tree

    if examples is empty then return default
    else if all examples have the same classification then return the classification
    else if attributes is empty then return MODE(examples)
    else
        best ← CHOOSE-ATTRIBUTE(attributes, examples)
        tree ← a new decision tree with root test best
        for each value $v_i$ of best do
            examples$_i$ ← \{elements of examples with best = $v_i$\}
            subtree ← DTL(examples$_i$, attributes - best, MODE(examples))
            add a branch to tree with label $v_i$ and subtree subtree
        return tree
```
Choosing an Attribute

- Idea: a good attribute splits the examples into subsets that are (ideally) “all positive” or “all negative”

- So: we need a measure of how “good” a split is, even if the results aren’t perfectly separated out
Entropy and Information

- **Information answers questions**
  - The more uncertain about the answer initially, the more information in the answer
  - Scale: bits
    - Answer to Boolean question with prior <1/2, 1/2>?
    - Answer to 4-way question with prior <1/4, 1/4, 1/4, 1/4>?
    - Answer to 4-way question with prior <0, 0, 0, 1>?
    - Answer to 3-way question with prior <1/2, 1/4, 1/4>?

- **A probability p is typical of:**
  - A uniform distribution of size 1/p
  - A code of length log 1/p
Entropy

- General answer: if prior is \( <p_1, \ldots, p_n> \):
  - Information is the expected code length

\[
H(\langle p_1, \ldots, p_n \rangle) = E_p \log_2 1/p_i
= \sum_{i=1}^{n} -p_i \log_2 p_i
\]

- Also called the entropy of the distribution
  - More uniform = higher entropy
  - More values = higher entropy
  - More peaked = lower entropy


1 bit

0 bits

0.5 bit
Information Gain

- Back to decision trees!
- For each split, compare entropy before and after
  - Difference is the information gain
  - Problem: there’s more than one distribution after split!

Solution: use expected entropy, weighted by the number of examples
Next Step: Recurse

- Now we need to keep growing the tree!
- Two branches are done (why?)
- What to do under “full”?  
  - See what examples are there...

<table>
<thead>
<tr>
<th>Example</th>
<th>Alt</th>
<th>Bar</th>
<th>Fri</th>
<th>Hun</th>
<th>Pat</th>
<th>Price</th>
<th>Rain</th>
<th>Res</th>
<th>Type</th>
<th>Est</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>X₁</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>Some</td>
<td>$$$</td>
<td>F</td>
<td>T</td>
<td>French</td>
<td>0–10</td>
<td>T</td>
</tr>
<tr>
<td>X₂</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>Full</td>
<td>$</td>
<td>F</td>
<td>F</td>
<td>Thai</td>
<td>30–60</td>
<td>F</td>
</tr>
<tr>
<td>X₃</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>Some</td>
<td>$</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>Burger</td>
<td>0–10</td>
</tr>
<tr>
<td>X₄</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>Full</td>
<td>$</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>Thai</td>
<td>10–30</td>
</tr>
<tr>
<td>X₅</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>Full</td>
<td>$$$</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>French</td>
<td>&gt;60</td>
</tr>
<tr>
<td>X₆</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>Some</td>
<td>$</td>
<td>T</td>
<td>T</td>
<td>Italian</td>
<td>0–10</td>
<td>T</td>
</tr>
<tr>
<td>X₇</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>None</td>
<td>$</td>
<td>T</td>
<td>F</td>
<td>Burger</td>
<td>0–10</td>
<td>F</td>
</tr>
<tr>
<td>X₈</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>Some</td>
<td>$</td>
<td>T</td>
<td>T</td>
<td>Thai</td>
<td>0–10</td>
<td>T</td>
</tr>
<tr>
<td>X₉</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>Full</td>
<td>$</td>
<td>T</td>
<td>F</td>
<td>Burger</td>
<td>&gt;60</td>
<td>F</td>
</tr>
<tr>
<td>X₁₀</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>Full</td>
<td>$$$</td>
<td>F</td>
<td>T</td>
<td>Italian</td>
<td>10–30</td>
<td>F</td>
</tr>
<tr>
<td>X₁₁</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>None</td>
<td>$</td>
<td>F</td>
<td>F</td>
<td>Thai</td>
<td>0–10</td>
<td>F</td>
</tr>
<tr>
<td>X₁₂</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>Full</td>
<td>$</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>Burger</td>
<td>30–60</td>
</tr>
</tbody>
</table>
Example: Learned Tree

- Decision tree learned from these 12 examples:

- Substantially simpler than “true” tree
  - A more complex hypothesis isn't justified by data
- Also: it’s reasonable, but wrong
Example: Miles Per Gallon

<table>
<thead>
<tr>
<th>mpg</th>
<th>cylinders</th>
<th>displacement</th>
<th>horsepower</th>
<th>weight</th>
<th>acceleration</th>
<th>modelyear</th>
<th>maker</th>
</tr>
</thead>
<tbody>
<tr>
<td>good</td>
<td>4</td>
<td>low</td>
<td>low</td>
<td>low</td>
<td>high</td>
<td>75to78</td>
<td>asia</td>
</tr>
<tr>
<td>bad</td>
<td>6</td>
<td>medium</td>
<td>medium</td>
<td>medium</td>
<td>medium</td>
<td>70to74</td>
<td>america</td>
</tr>
<tr>
<td>bad</td>
<td>4</td>
<td>medium</td>
<td>medium</td>
<td>low</td>
<td>75to78</td>
<td>70to74</td>
<td>america</td>
</tr>
<tr>
<td>bad</td>
<td>8</td>
<td>high</td>
<td>high</td>
<td>low</td>
<td>70to74</td>
<td>70to74</td>
<td>america</td>
</tr>
<tr>
<td>bad</td>
<td>6</td>
<td>medium</td>
<td>medium</td>
<td>medium</td>
<td>70to74</td>
<td>70to74</td>
<td>america</td>
</tr>
<tr>
<td>bad</td>
<td>4</td>
<td>low</td>
<td>medium</td>
<td>medium</td>
<td>70to74</td>
<td>70to74</td>
<td>asia</td>
</tr>
<tr>
<td>bad</td>
<td>4</td>
<td>low</td>
<td>medium</td>
<td>medium</td>
<td>70to74</td>
<td>70to74</td>
<td>asia</td>
</tr>
<tr>
<td>bad</td>
<td>8</td>
<td>high</td>
<td>high</td>
<td>low</td>
<td>75to78</td>
<td>america</td>
<td></td>
</tr>
<tr>
<td>good</td>
<td>8</td>
<td>high</td>
<td>high</td>
<td>low</td>
<td>70to74</td>
<td>america</td>
<td></td>
</tr>
<tr>
<td>good</td>
<td>8</td>
<td>high</td>
<td>medium</td>
<td>high</td>
<td>high</td>
<td>79to83</td>
<td>america</td>
</tr>
<tr>
<td>bad</td>
<td>8</td>
<td>high</td>
<td>high</td>
<td>low</td>
<td>75to78</td>
<td>america</td>
<td></td>
</tr>
<tr>
<td>good</td>
<td>4</td>
<td>low</td>
<td>low</td>
<td>low</td>
<td>low</td>
<td>79to83</td>
<td>america</td>
</tr>
<tr>
<td>bad</td>
<td>6</td>
<td>medium</td>
<td>medium</td>
<td>medium</td>
<td>high</td>
<td>75to78</td>
<td>america</td>
</tr>
<tr>
<td>good</td>
<td>4</td>
<td>medium</td>
<td>low</td>
<td>low</td>
<td>low</td>
<td>79to83</td>
<td>america</td>
</tr>
<tr>
<td>good</td>
<td>4</td>
<td>low</td>
<td>medium</td>
<td>high</td>
<td>79to83</td>
<td>america</td>
<td></td>
</tr>
<tr>
<td>bad</td>
<td>8</td>
<td>high</td>
<td>high</td>
<td>low</td>
<td>70to74</td>
<td>america</td>
<td></td>
</tr>
<tr>
<td>good</td>
<td>4</td>
<td>low</td>
<td>medium</td>
<td>low</td>
<td>medium</td>
<td>75to78</td>
<td>europe</td>
</tr>
<tr>
<td>bad</td>
<td>5</td>
<td>medium</td>
<td>medium</td>
<td>medium</td>
<td>75to78</td>
<td>europe</td>
<td></td>
</tr>
</tbody>
</table>
Find the First Split

- Look at information gain for each attribute

- Note that each attribute is correlated with the target!

- What do we split on?
Result: Decision Stump

mpg values: bad good

root
22 18
pchance = 0.001

cylinders = 3
0 0
Predict bad

cylinders = 4
4 17
Predict good

cylinders = 5
1 0
Predict bad

cylinders = 6
8 0
Predict bad

cylinders = 8
9 1
Predict bad
mpg values: bad good

root
22 18
pchance = 0.001

cylinders = 3
0 0
Predict bad

0 1
Predict good

0 1
Predict bad

cylinders = 4
4 17
pchance = 0.135

0 1
Predict bad

0 1
Predict bad

cylinders = 5
1 0
pchance = 0.085

cylinders = 6
8 0

pchance = 0.085

pchance = 0.085

cylinders = 8
9 1

pchance = 0.085

maker = america
0 10
Predict good

0 10
Predict good

0 10
Predict good

maker = asia
2 5
Predict good

2 5
Predict good

2 5
Predict good

maker = europe
2 2
Predict bad

2 2
Predict bad

2 2
Predict bad

horsepower = low
0 0

0 0

0 0

horsepower = medium
0 1

0 1

0 1

horsepower = high
9 0

9 0

9 0

9 0

pchance = 0.085

pchance = 0.085

pchance = 0.085
Reminder: Overfitting

- **Overfitting:**
  - When you stop modeling the patterns in the training data (which generalize)
  - And start modeling the noise (which doesn’t)

- **We had this before:**
  - Naïve Bayes: needed to smooth
  - Perceptron: early stopping
The test set error is much worse than the training set error...

...why?
Significance of a Split

- **Starting with:**
  - Three cars with 4 cylinders, from Asia, with medium HP
  - 2 bad MPG
  - 1 good MPG

- **What do we expect from a three-way split?**
  - Maybe each example in its own subset?
  - Maybe just what we saw in the last slide?

- Probably shouldn’t split if the counts are so small they could be due to chance

- A chi-squared test can tell us how likely it is that deviations from a perfect split are due to chance

- Each split will have a *significance value, p*$_{\text{CHANCE}}$
Pruning:
- Build the full decision tree
- Begin at the bottom of the tree
- Delete splits in which $p_{\text{CHANCE}} > \text{MaxP}_{\text{CHANCE}}$
- Continue working upward until there are no more prunable nodes

$y = a \text{ XOR } b$

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Pruning example

- With $\text{MaxP}_{\text{CHANCE}} = 0.1$:

Note the improved test set accuracy compared with the unpruned tree.

<table>
<thead>
<tr>
<th></th>
<th>Num Errors</th>
<th>Set Size</th>
<th>Percent Wrong</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Set</td>
<td>5</td>
<td>40</td>
<td>12.50</td>
</tr>
<tr>
<td>Test Set</td>
<td>56</td>
<td>352</td>
<td>15.91</td>
</tr>
</tbody>
</table>
- MaxP\textsubscript{CHANCE} is a regularization parameter
- Generally, set it using held-out data (as usual)
Forward Pointer: Random Forests

- Ensemble method of learning trees
- Very effective, used in industry and more
- Related to Boosting
A few important points about learning

- **Data:** labeled instances, e.g. emails marked spam/ham
  - Training set
  - Held out set
  - Test set
- **Features:** attribute-value pairs which characterize each x
- **Experimentation cycle**
  - Learn parameters (e.g. model probabilities) on training set
  - (Tune hyperparameters on held-out set)
  - Compute accuracy of test set
  - Very important: never “peek” at the test set!
- **Evaluation**
  - Accuracy: fraction of instances predicted correctly
- **Overfitting and generalization**
  - Want a classifier which does well on test data
  - Overfitting: fitting the training data very closely, but not generalizing well
  - Underfitting: fits the training set poorly
A few important points about learning

- **What should we learn where?**
  - Learn parameters from training data
  - Tune hyperparameters on different data
    - Why?
  - For each value of the hyperparameters, train and test on the held-out data
  - Choose the best value and do a final test on the test data

- **What are examples of hyperparameters?**
Inductive Learning
Inductive Learning (Science)

- Simplest form: learn a function from examples
  - A target function: $g$
  - Examples: input-output pairs $(x, g(x))$
  - E.g. $x$ is an email and $g(x)$ is spam / ham
  - E.g. $x$ is a house and $g(x)$ is its selling price

- Problem:
  - Given a hypothesis space $H$
  - Given a training set of examples $x_i$
  - Find a hypothesis $h(x)$ such that $h \sim g$

- Includes:
  - Classification (outputs = class labels)
  - Regression (outputs = real numbers)

- How do perceptron and naïve Bayes fit in? ($H, h, g$, etc.)
Inductive Learning

- Curve fitting (regression, function approximation):
  - Consistency vs. simplicity
  - Ockham’s razor
Consistency vs. Simplicity

- Fundamental tradeoff: bias vs. variance
- Usually algorithms prefer consistency by default (why?)

Several ways to operationalize “simplicity”
- Reduce the hypothesis space
  - Assume more: e.g. independence assumptions, as in naïve Bayes
  - Have fewer, better features / attributes: feature selection
  - Other structural limitations (decision lists vs trees)
- Regularization
  - Smoothing: cautious use of small counts
  - Many other generalization parameters (pruning cutoffs today)
  - Hypothesis space stays big, but harder to get to the outskirts