Neural Nets (wrap-up) and Decision Trees

Today

- Neural Nets -- wrap
- Formalizing Learning
  - Consistency
  - Simplicity
- Decision Trees
  - Expressiveness
  - Information Gain
  - Overfitting

Deep Neural Network

Deep Neural Network: Also Learn the Features!

- Training the deep neural network is just like logistic regression:
  \[
  \max_w \log \mathbb{P}(y^{(i)} | x^{(i)}; w) = \sum_i \log \mathbb{P}(y^{(i)} | x^{(i)}; w)
  \]
  
  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 well does it work?
Performance

ImageNet Error Rate 2010-2014

Graph credit Matt Zeiler, Clarifai

AlexNet Performance

Graph credit Matt Zeiler, Clarifai

MS COCO Image Captioning Challenge

Karpathy & Fei-Fei, 2015; Donahue et al., 2015; Xu et al, 2015; many more

Visual QA Challenge

Standford Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Chris Barts, C. Lawrence Zitnick, Devi Parikh
Today

- Neural Nets -- wrap
- Formalizing Learning
  - Consistency
  - Simplicity
- Decision Trees
  - Expressiveness
  - Information Gain
  - Overfitting
- Clustering

Inductive Learning

- Simplest form: learn a function from examples
- A target function: \( g \)
- Examples: input-output pairs: \( (x_i, g(x_i)) \)
- 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 \( X \)
  - Find a hypothesis \( h(x) \) such that \( h \sim g \)

- Includes:
  - Classification (outputs = class labels)
  - Regression (outputs = real numbers)
  - How doesperceptron and naïve Bayes fit in? \( (H, \sim, g, \text{etc.}) \)

Inductive Learning (Science)

- Curve fitting (regression, function approximation):

Inductive Learning

- 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

**Decision Trees**

- Compact representation of a function:
  - Truth table
  - Conditional probability table
  - Regression values
- True function
  - Realizable: in \( \mathcal{H} \)

**Expressiveness of DTs**

- Can express any function of the features
- However, we hope for compact trees

**Decision 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 Ù 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
Hypothesis Spaces

- How many distinct decision trees with \( n \) Boolean attributes?
  - \# number of Boolean functions over \( n \) attributes
  - \# number of distinct truth tables with \( 2^n \) rows
  - \( \sum_{n=1}^{2^n} \)
  - E.g., with 6 Boolean attributes, there are 18,446,744,073,709,551,616 trees.

- How many trees of depth 1 (decision stumps)?
  - \# number of Boolean functions over 1 attribute
  - \# number of truth tables with 2 rows, times \( n \)
  - \( 4^n \)
  - E.g., with 6 Boolean attributes, there are 24 decision stumps.

- More expressive hypothesis space:
  - Increases chance that target function can be expressed (good)
  - Increases number of hypotheses consistent with training set (bad, why?)
  - Means we can get better predictions (lower bias)
  - But we may get worse predictions (higher variance)

Decision Tree Learning

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

```
function DTL(examples, attributes, depth)
  if examples is empty then return default
  else if all examples have the same classification then return the classification
  else if attribute is empty then return Most(examples)
  else
    best -- CHOOSE-ATTRIBUTE(attributes, examples)
    tree -- a new decision tree with root test best
    for each value \( v \) of best do
      examples -- (elements of examples with best = \( v \))
      subtree -- DTL(examples, attributes -- best, Most(examples))
      add a branch to tree with label \( v \) 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>?
  - 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
  
  \[
  H(p_1, \ldots, p_n) = E_{\mathbf{p}} \log_2 1/p_i = \sum_{i=1}^{n} -p_i \log_2 p_i
  \]

  - Also called the entropy of the distribution
    - More uniform = higher entropy
    - More values = higher entropy
    - More peaked = lower entropy
    - Rare values almost “don’t count”

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...

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

- 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
  - 23 10
  - purity = 0.001

  cylinders = 3
cylinders = 4
cylinders = 5
cylinders = 6
cylinders = 8

  Predict bad Predict good Predict bad Predict bad Predict bad

Second Level

- mpg values: bad good
  - root
  - 22 10
  - purity = 0.001

  cylinders > 5
cylinders = 5
cylinders = 4
cylinders = 3

  Predict bad Predict bad Predict bad Predict bad Predict bad

  cylinders = 2
cylinders = 1
cylinders >= 0

  Predict bad Predict bad Predict bad Predict bad Predict bad

  Predicted mpg range [0.001, 12.000]
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

### 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**, $p_{\text{CHANCE}}$

### Keeping it General

- **Pruning:**
  - Build the full decision tree
  - Begin at the bottom of the tree
  - Delete splits in which $p_{\text{CHANCE}} > \text{Max} p_{\text{CHANCE}}$
  - Continue working upward until there are no more prunable nodes

- Note: some chance nodes may not get pruned because they were “redeemed” later

### MPG Training Error

- The test set error is much worse than the training set error...
  - ...why?
Pruning example

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

<table>
<thead>
<tr>
<th>Error Train</th>
<th>Test</th>
<th>Val</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train 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>

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

Regularization

- $\text{MaxP}_{\text{CHANCE}}$ is a regularization parameter
- Generally, set it using held-out data (as usual)

Two Ways of Controlling Overfitting

- **Limit the hypothesis space**
  - E.g. limit the max depth of trees
  - Easier to analyze

- **Regularize the hypothesis selection**
  - E.g. chance cutoff
  - Disprefer most of the hypotheses unless data is clear
  - Usually done in practice

Next Lecture: Applications!