Introduction & Model

Introduction to Machine Learning & Models

Poul Petersen
CIO, BigML, Inc
A Brief History of BigML

- BigML Mission: To make Machine Learning **Beautifully Simple**
- BigML Founded in **Corvallis**, Oregon in 2011 - long before ML was "cool"
- You’ve never heard of it?
- Most innovative city in the United States!
Statistics show Corvallis No. 1 for patents

By BENNETT HALL, Gazette-Times reporter  Jan 25, 2011  0
BigML Platform

Visualizations

Web-based Frontend

Tools - https://bigml.com/tools

REST API - https://bigml.com/api

Distributed Machine Learning Backend

Smart Infrastructure (auto-deployable, auto-scalable)
Who am "I"?

- Poul Petersen, BigML, CIO since Nov 2011
- Background in Mathematics, Physics, Engineering
- Wrote "Sauron"
  - Fully automated bigml.com, bigml.com.au
  - Ported Sauron for private deployments
- Created:
  - BigML for Alexa (& house recommender)
  - Original PredictServer
  - First "scriptify", called "reifier"
- Demos, Training, and Sales Support
Who are you?

**Expert:** Published papers at KDD, ICML, NIPS, etc or developed own ML algorithms used at large scale

**Aficionado:** Understands pros/cons of different techniques and/or can tweak algorithms as needed

**Practitioner:** Very familiar with ML packages (Weka, Scikit, BigML, etc.)

**Newbie:** Just taking Coursera ML class or reading an introductory book to ML

**Absolute beginner:** ML sounds like science fiction
Before we Begin…

Pacing Goal

Reality
Before we Begin…

- All course materials including slides, CSVs and related videos are hosted on the Training website:
  
  http://training.bigml.com

- Each module will have learning exercises to be performed in-session, so…

- You need a laptop and a BigML account: Get signed up

  https://bigml.com/account/register
Machine Learning Motivation

Imagine:

- You are looking to buy a house
- Recently found a house you like
- Is the asking price fair?

What Next?
Machine Learning Motivation

Why not ask an expert?

• Experts can be rare / expensive
• Hard to validate experience:
  • Experience with similar properties?
  • Do they consider all relevant variables?
  • Knowledge of market up to date?
• Hard to validate answer:
  • How many times expert right / wrong?
  • Probably can’t explain decision in detail
• Humans are not good at intuitive statistics
Data vs Expert

Replace the expert with data?

• Intuition: square footage relates to price.
• Collect data from past sales

<table>
<thead>
<tr>
<th>SQFT</th>
<th>SOLD</th>
<th>PREDICT</th>
</tr>
</thead>
<tbody>
<tr>
<td>2424</td>
<td>360000</td>
<td>400262</td>
</tr>
<tr>
<td>1785</td>
<td>307500</td>
<td>320195</td>
</tr>
<tr>
<td>1003</td>
<td>185000</td>
<td>222211</td>
</tr>
<tr>
<td>4135</td>
<td>600000</td>
<td>614651</td>
</tr>
<tr>
<td>1676</td>
<td>328500</td>
<td>306538</td>
</tr>
<tr>
<td>1012</td>
<td>247000</td>
<td>223339</td>
</tr>
<tr>
<td>3352</td>
<td>420000</td>
<td>516541</td>
</tr>
<tr>
<td>2825</td>
<td>435350</td>
<td>450508</td>
</tr>
</tbody>
</table>

PRICE = 125.3*SQFT + 96535
Data vs Expert

Replace the expert scorecard

- Experts can be rare / expensive
- Hard to validate experience:
  - Experience with similar properties?
  - Do they consider all relevant variables?
  - Knowledge of market up to date?
- Hard to validate answer:
  - How many times expert right / wrong?
  - Probably can’t explain decision in detail
- Humans are not good at intuitive statistics
Replace the expert with data

- Intuition: square footage relates to price.
- Collect data from past sales

<table>
<thead>
<tr>
<th>SQFT</th>
<th>SOLD</th>
</tr>
</thead>
<tbody>
<tr>
<td>2424</td>
<td>360000</td>
</tr>
<tr>
<td>1785</td>
<td>307500</td>
</tr>
<tr>
<td>1003</td>
<td>185000</td>
</tr>
<tr>
<td>4135</td>
<td>600000</td>
</tr>
<tr>
<td>1676</td>
<td>328500</td>
</tr>
<tr>
<td>1012</td>
<td>247000</td>
</tr>
<tr>
<td>3352</td>
<td>420000</td>
</tr>
<tr>
<td>2825</td>
<td>435350</td>
</tr>
</tbody>
</table>

PRICE = 125.3*SQFT + 96535
More Data!

<table>
<thead>
<tr>
<th>SQFT</th>
<th>BEDS</th>
<th>BATHS</th>
<th>ADDRESS</th>
<th>LOCATION</th>
<th>LOT SIZE</th>
<th>YEAR BUILT</th>
<th>PARKING SPOTS</th>
<th>LATITUDE</th>
<th>LONGITUDE</th>
<th>SOLD</th>
</tr>
</thead>
<tbody>
<tr>
<td>2424</td>
<td>4</td>
<td>3</td>
<td>1522 NW Jonquil</td>
<td>Timberhill SE 2nd</td>
<td>5227</td>
<td>1991</td>
<td>2</td>
<td>44.594828</td>
<td>-123.269328</td>
<td>360000</td>
</tr>
<tr>
<td>1785</td>
<td>3</td>
<td>2</td>
<td>7360 NW Valley Vw</td>
<td>Country Estates</td>
<td>25700</td>
<td>1979</td>
<td>2</td>
<td>44.643876</td>
<td>-123.238189</td>
<td>307500</td>
</tr>
<tr>
<td>1003</td>
<td>2</td>
<td>1</td>
<td>2620 NW Chinaberry</td>
<td>Tamarack Village</td>
<td>4792</td>
<td>1978</td>
<td>2</td>
<td>44.593704</td>
<td>-123.295424</td>
<td>185000</td>
</tr>
<tr>
<td>4135</td>
<td>5</td>
<td>3.5</td>
<td>4748 NW Veronica</td>
<td>Suncrest</td>
<td>6098</td>
<td>2004</td>
<td>3</td>
<td>44.5929659</td>
<td>-123.306916</td>
<td>600000</td>
</tr>
<tr>
<td>1676</td>
<td>3</td>
<td>2</td>
<td>2842 NW Monterey</td>
<td>Corvallis</td>
<td>8712</td>
<td>1975</td>
<td>2</td>
<td>44.5945279</td>
<td>-123.291523</td>
<td>328500</td>
</tr>
<tr>
<td>1012</td>
<td>3</td>
<td>1</td>
<td>2320 NW Highland</td>
<td>Corvallis</td>
<td>9583</td>
<td>1959</td>
<td>2</td>
<td>44.591476</td>
<td>-123.262841</td>
<td>247000</td>
</tr>
<tr>
<td>3352</td>
<td>4</td>
<td>3</td>
<td>1205 NW Ridgewood</td>
<td>Ridgewood 2</td>
<td>60113</td>
<td>1975</td>
<td>2</td>
<td>44.579439</td>
<td>-123.333888</td>
<td>420000</td>
</tr>
<tr>
<td>2825</td>
<td>3</td>
<td></td>
<td>411 NW 16th</td>
<td>Wilkins Addition</td>
<td>4792</td>
<td>1938</td>
<td>1</td>
<td>44.570883</td>
<td>-123.272113</td>
<td>435350</td>
</tr>
</tbody>
</table>

Uhhhh……

- Can we still fit a line to 10 variables? (well, yes)
- Will fitting a line give good results? (unlikely)
- What about those text fields and categorical values?
Linear Regression
What Just Happened?

- We started with Housing data as a CSV from Redfin
- We uploaded the CSV to create Source
- Then we created a Dataset from the Source and reviewed the summary statistics and scatter plot
- With 1-click we build a Linear Regression which can predict home prices based on all the housing features
- We explored the Model and used it to make a Prediction of a home with SQFT=1000 and LOT SIZE=0
- We noticed that it was impossible to disable some fields like LOCATION in the prediction form.
Aside: BigML Resources

- Everything created on BigML is a resource
- Resources are:
  - Immutable: Ensures consistency, repeatability
  - Assigned a canonical ID
  - Traceable: How the resource is created is part of resource
  - Always available via both the API and the UI
  - Sharable with a secret link, api key, or organization
- Working with BigML is a process of creating resources
- Resources can be grouped into Projects
  - Projects can be private or part of an organization
  - Organizations allow real-time sharing of all resources in a project
Your Turn!

• Create a new project called “Training”
• Build a Linear Regression for the PDX dataset
  • The CSV is in http://training.bigml.com
• What is the predicted price of:
  • SQFT = 1,000, LOT SIZE=0
  • All other fields at default
  • What inputs can you not disable?
• Bonus:
  • Share your model with someone…
More Data!

<table>
<thead>
<tr>
<th>SQFT</th>
<th>BEDS</th>
<th>BATHS</th>
<th>ADDRESS</th>
<th>LOCATION</th>
<th>LOT SIZE</th>
<th>YEAR BUILT</th>
<th>PARKING SPOTS</th>
<th>LATITUDE</th>
<th>LONGITUDE</th>
<th>SOLD</th>
</tr>
</thead>
<tbody>
<tr>
<td>2424</td>
<td>4</td>
<td>3</td>
<td>1522 NW Jonquil</td>
<td>Timberhill SE 2nd</td>
<td>5227</td>
<td>1991</td>
<td>2</td>
<td>44.594828</td>
<td>-123.269328</td>
<td>360000</td>
</tr>
<tr>
<td>1785</td>
<td>3</td>
<td>2</td>
<td>7360 NW Valley Vw</td>
<td>Country Estates</td>
<td>25700</td>
<td>1979</td>
<td>2</td>
<td>44.643876</td>
<td>-123.238189</td>
<td>307500</td>
</tr>
<tr>
<td>1003</td>
<td>2</td>
<td>1</td>
<td>2620 NW Chinaberry</td>
<td>Tamarack Village</td>
<td>4792</td>
<td>1978</td>
<td>2</td>
<td>44.593704</td>
<td>-123.295424</td>
<td>185000</td>
</tr>
<tr>
<td>4135</td>
<td>5</td>
<td>3.5</td>
<td>4748 NW Veronica</td>
<td>Suncrest</td>
<td>6098</td>
<td>2004</td>
<td>3</td>
<td>44.5929659</td>
<td>-123.306916</td>
<td>600000</td>
</tr>
<tr>
<td>1676</td>
<td>3</td>
<td>2</td>
<td>2842 NW Monterey</td>
<td>Corvallis</td>
<td>8712</td>
<td>1975</td>
<td>2</td>
<td>44.5945279</td>
<td>-123.291523</td>
<td>328500</td>
</tr>
<tr>
<td>1012</td>
<td>3</td>
<td>1</td>
<td>2320 NW Highland</td>
<td>Corvallis</td>
<td>9583</td>
<td>1959</td>
<td>2</td>
<td>44.591476</td>
<td>-123.262841</td>
<td>247000</td>
</tr>
<tr>
<td>3352</td>
<td>4</td>
<td>3</td>
<td>1205 NW Ridgewood</td>
<td>Ridgewood 2</td>
<td>60113</td>
<td>1975</td>
<td>2</td>
<td>44.579439</td>
<td>-123.333888</td>
<td>420000</td>
</tr>
<tr>
<td>2825</td>
<td>3</td>
<td></td>
<td>411 NW 16th</td>
<td>Wilkins Addition</td>
<td>4792</td>
<td>1938</td>
<td>1</td>
<td>44.570883</td>
<td>-123.272113</td>
<td>435350</td>
</tr>
</tbody>
</table>

Uhhhh……..

• Can we still fit a line to 10 variables? (well, yes)
• Will fitting a line give good results? (unlikely)
• What about those text fields and categorical values?
Mythical ML Model?

- High representational power
  - Fitting a line is an example of low
  - Deep neural networks is an example of high
- High Ease-of-use
  - Easy to configure - relatively few parameters
  - Easy to interpret - how are decisions made?
  - Easy to put into production
- Ability to work with real-world data
  - Mixed data types: numeric, categorical, text, etc
  - Handle missing values
  - Resilient to outliers
- There are actually hundreds of possible choices…
Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?

Manuel Fernández-Delgado
Eva Cernadas
Senén Barro
CITIUS: Centro de Investigación en Tecnologías da Información da USC
University of Santiago de Compostela
Campus Vida, 15872, Santiago de Compostela, Spain

Dinani Amorim
Departamento de Tecnologia e Ciências Sociais- DTCS
Universidade do Estado da Bahia
Av. Edgard Chastinet S/N - São Geraldo - Juazeiro-BA, CEP: 48.305-680, Brasil

Editor: Russ Greiner

Abstract

We evaluate 179 classifiers arising from 17 families (discriminant analysis, Bayesian, neural networks, support vector machines, decision trees, rule-based classifiers, boosting, bagging, stacking, random forests and other ensembles, generalized linear models, nearest-neighbors, partial least squares and principal component regression, logistic and multinomial regression, multiple adaptive regression splines and other methods), implemented in Weka, R (with and without the caret package), C and Matlab, including all the relevant classifiers available today. We use 121 data sets, which represent the whole UCI data base (excluding the large-scale problems) and other own real problems, in order to achieve significant conclusions about the classifier behavior, not dependent on the data set collection. The classifiers most likely to be the bests are the random forest (RF) variant, the best of which (implemented in R and accessed via caret) achieves 94.1% of
A churn problem...

<table>
<thead>
<tr>
<th>Minutes Used</th>
<th>Last Month’s Bill</th>
<th>Calls To Support</th>
<th>Website Visits</th>
<th>Churn?</th>
</tr>
</thead>
<tbody>
<tr>
<td>104</td>
<td>103.60</td>
<td>0</td>
<td>0</td>
<td>No</td>
</tr>
<tr>
<td>124</td>
<td>56.33</td>
<td>1</td>
<td>0</td>
<td>No</td>
</tr>
<tr>
<td>56</td>
<td>214.60</td>
<td>2</td>
<td>0</td>
<td>Yes</td>
</tr>
<tr>
<td>2410</td>
<td>305.60</td>
<td>0</td>
<td>5</td>
<td>No</td>
</tr>
<tr>
<td>536</td>
<td>145.70</td>
<td>0</td>
<td>0</td>
<td>No</td>
</tr>
<tr>
<td>234</td>
<td>122.09</td>
<td>0</td>
<td>1</td>
<td>No</td>
</tr>
<tr>
<td>201</td>
<td>185.76</td>
<td>1</td>
<td>7</td>
<td>Yes</td>
</tr>
<tr>
<td>111</td>
<td>83.60</td>
<td>3</td>
<td>2</td>
<td>No</td>
</tr>
</tbody>
</table>

Hey! Those are not numbers!
Supervised Learning

### Regression

<table>
<thead>
<tr>
<th>animal</th>
<th>state</th>
<th>proximity</th>
<th>min_kmph</th>
</tr>
</thead>
<tbody>
<tr>
<td>tiger</td>
<td>hungry</td>
<td>close</td>
<td>70</td>
</tr>
<tr>
<td>hippo</td>
<td>angry</td>
<td>far</td>
<td>10</td>
</tr>
</tbody>
</table>

### Classification

<table>
<thead>
<tr>
<th>animal</th>
<th>state</th>
<th>proximity</th>
<th>action</th>
</tr>
</thead>
<tbody>
<tr>
<td>tiger</td>
<td>hungry</td>
<td>close</td>
<td>run</td>
</tr>
<tr>
<td>elephant</td>
<td>happy</td>
<td>far</td>
<td>take picture</td>
</tr>
</tbody>
</table>

### Multi-Label Classification

<table>
<thead>
<tr>
<th>animal</th>
<th>state</th>
<th>proximity</th>
<th>action1</th>
<th>action2</th>
</tr>
</thead>
<tbody>
<tr>
<td>tiger</td>
<td>hungry</td>
<td>close</td>
<td>run</td>
<td>look untasty</td>
</tr>
<tr>
<td>elephant</td>
<td>happy</td>
<td>far</td>
<td>take picture</td>
<td>call friends</td>
</tr>
</tbody>
</table>

Question: What would Unsupervised Learning look like?
ML: Two Methods

**Supervised**
- Requires labelled data
- Goal is to predict the label, often called the objective
- Can be evaluated against the label
- Algorithms:
  - Models/Ensembles
  - Logistic Regression
  - Deepnets
  - Time Series

**Unsupervised**
- Does not require labelled data
- Goal is “discovery”, with algorithms focused on type
- Each algorithm has its own quality measures
- Algorithms:
  - Clustering
  - Anomaly Detection
  - Association Discovery
  - Topic Modeling
## Back to the Data…

<table>
<thead>
<tr>
<th>Minutes Used</th>
<th>Last Month’s Bill</th>
<th>Calls To Support</th>
<th>Website Visits</th>
<th>Churn?</th>
</tr>
</thead>
<tbody>
<tr>
<td>104</td>
<td>103.60</td>
<td>0</td>
<td>0</td>
<td>No</td>
</tr>
<tr>
<td>124</td>
<td>56.33</td>
<td>1</td>
<td>0</td>
<td>No</td>
</tr>
<tr>
<td>56</td>
<td>214.60</td>
<td>2</td>
<td>0</td>
<td>Yes</td>
</tr>
<tr>
<td>2410</td>
<td>305.60</td>
<td>0</td>
<td>5</td>
<td>No</td>
</tr>
<tr>
<td>536</td>
<td>145.70</td>
<td>0</td>
<td>0</td>
<td>No</td>
</tr>
<tr>
<td>234</td>
<td>122.09</td>
<td>0</td>
<td>1</td>
<td>No</td>
</tr>
<tr>
<td>201</td>
<td>185.76</td>
<td>1</td>
<td>7</td>
<td>Yes</td>
</tr>
<tr>
<td>111</td>
<td>83.60</td>
<td>3</td>
<td>2</td>
<td>No</td>
</tr>
<tr>
<td>Minutes Used</td>
<td>Last Month’s Bill</td>
<td>Calls To Support</td>
<td>Website Visits</td>
<td>Churn?</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------------</td>
<td>------------------</td>
<td>----------------</td>
<td>-------</td>
</tr>
<tr>
<td>104</td>
<td>103.60</td>
<td>0</td>
<td>0</td>
<td>No</td>
</tr>
<tr>
<td>124</td>
<td>56.33</td>
<td>1</td>
<td>0</td>
<td>No</td>
</tr>
<tr>
<td>56</td>
<td>214.60</td>
<td>2</td>
<td>0</td>
<td>Yes</td>
</tr>
<tr>
<td>2410</td>
<td>305.60</td>
<td>0</td>
<td>5</td>
<td>No</td>
</tr>
<tr>
<td>536</td>
<td>145.70</td>
<td>0</td>
<td>0</td>
<td>No</td>
</tr>
<tr>
<td>234</td>
<td>122.09</td>
<td>0</td>
<td>1</td>
<td>No</td>
</tr>
<tr>
<td>201</td>
<td>185.76</td>
<td>1</td>
<td>7</td>
<td>Yes</td>
</tr>
<tr>
<td>111</td>
<td>83.60</td>
<td>3</td>
<td>2</td>
<td>No</td>
</tr>
<tr>
<td>Minutes Used</td>
<td>Last Month’s Bill</td>
<td>Calls To Support</td>
<td>Website Visits</td>
<td>Churn?</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------------</td>
<td>------------------</td>
<td>----------------</td>
<td>--------</td>
</tr>
<tr>
<td>104</td>
<td>103.60</td>
<td>0</td>
<td>0</td>
<td>No</td>
</tr>
<tr>
<td>124</td>
<td>56.33</td>
<td>1</td>
<td>0</td>
<td>No</td>
</tr>
<tr>
<td>56</td>
<td>214.60</td>
<td>2</td>
<td>0</td>
<td>Yes</td>
</tr>
<tr>
<td>2410</td>
<td>305.60</td>
<td>0</td>
<td>5</td>
<td>No</td>
</tr>
<tr>
<td>536</td>
<td>145.70</td>
<td>0</td>
<td>0</td>
<td>No</td>
</tr>
<tr>
<td>234</td>
<td>122.09</td>
<td>0</td>
<td>1</td>
<td>No</td>
</tr>
<tr>
<td>201</td>
<td>185.76</td>
<td>1</td>
<td>7</td>
<td>Yes</td>
</tr>
<tr>
<td>111</td>
<td>83.60</td>
<td>3</td>
<td>2</td>
<td>No</td>
</tr>
<tr>
<td>Minutes Used</td>
<td>Last Month’s Bill</td>
<td>Calls To Support</td>
<td>Website Visits</td>
<td>Churn?</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------------</td>
<td>------------------</td>
<td>----------------</td>
<td>--------</td>
</tr>
<tr>
<td>104</td>
<td>103.60</td>
<td>0</td>
<td>0</td>
<td>No</td>
</tr>
<tr>
<td>124</td>
<td>56.33</td>
<td>1</td>
<td>0</td>
<td>No</td>
</tr>
<tr>
<td>56</td>
<td>214.60</td>
<td>2</td>
<td>0</td>
<td>Yes</td>
</tr>
<tr>
<td>2410</td>
<td>305.60</td>
<td>0</td>
<td>5</td>
<td>No</td>
</tr>
<tr>
<td>536</td>
<td>145.70</td>
<td>0</td>
<td>0</td>
<td>No</td>
</tr>
<tr>
<td>234</td>
<td>122.09</td>
<td>0</td>
<td>1</td>
<td>No</td>
</tr>
<tr>
<td>201</td>
<td>185.76</td>
<td>1</td>
<td>7</td>
<td>Yes</td>
</tr>
<tr>
<td>111</td>
<td>83.60</td>
<td>3</td>
<td>2</td>
<td>No</td>
</tr>
</tbody>
</table>
## Last Bill > $180, Support Calls > 0

<table>
<thead>
<tr>
<th>Minutes Used</th>
<th>Last Month’s Bill</th>
<th>Calls To Support</th>
<th>Website Visits</th>
<th>Churn?</th>
</tr>
</thead>
<tbody>
<tr>
<td>104</td>
<td>103.60</td>
<td>0</td>
<td>0</td>
<td>No</td>
</tr>
<tr>
<td>124</td>
<td>56.33</td>
<td>1</td>
<td>0</td>
<td>No</td>
</tr>
<tr>
<td>56</td>
<td>214.60</td>
<td>2</td>
<td>0</td>
<td>Yes</td>
</tr>
<tr>
<td>2410</td>
<td>305.60</td>
<td>0</td>
<td>5</td>
<td>No</td>
</tr>
<tr>
<td>536</td>
<td>145.70</td>
<td>0</td>
<td>0</td>
<td>No</td>
</tr>
<tr>
<td>234</td>
<td>122.09</td>
<td>0</td>
<td>1</td>
<td>No</td>
</tr>
<tr>
<td>201</td>
<td>185.76</td>
<td>1</td>
<td>7</td>
<td>Yes</td>
</tr>
<tr>
<td>111</td>
<td>83.60</td>
<td>3</td>
<td>2</td>
<td>No</td>
</tr>
</tbody>
</table>
Models
What Just Happened?

- We started with Churn data as a CSV
- We uploaded the CSV to create **Source**
- Then we created a **Dataset** from the **Source** and reviewed the summary statistics
- With 1-click we build a **Model** which can predict which customers will churn.
- We explored the **Model** and used it to make a **Prediction**
Why Decision Trees

- Works for classification or regression
- Easy to understand: splits are features and values
- Lightweight and super fast at prediction time
DT Predictions

Question 1

Question 2

Prediction
Why Decision Trees

- Works for classification or regression
- Easy to understand: splits are features and values
- Lightweight and super fast at prediction time
- Relatively parameter free
- Data can be messy
  - Useless features are automatically ignored
  - Works with un-normalized data
  - Works with missing data at Training
Training with Missing

Loan Amount?

Reason: Missing?
Why Decision Trees

- Works for classification or regression
- Easy to understand: splits are features and values
- Lightweight and super fast at prediction time
- Relatively parameter free
- Data can be messy
  - Useless features are automatically ignored
  - Works with un-normalized data
  - Works with missing data at Training & Prediction
Predictions with Missing

Question 1

Missing?

Last Prediction
Predictions with Missing

- Question 1
- Missing?
- Question 2
- Avg Prediction

Skip

Question 3
Why Decision Trees

- Works for classification or regression
- Easy to understand: splits are features and values
- Lightweight and super fast at prediction time
- Relatively parameter free
- Data can be messy
  - Useless features are automatically ignored
  - Works with un-normalized data
  - Works with missing data at Training & Prediction
  - Resilient to outliers
- High representational power
- Works easily with mixed data types
Why Not Decision Trees

- Slightly prone to over-fitting. (wait: what is that?)
Learning Problems (fit)

Under-fitting

- Model does not fit well enough
- Does not capture the underlying trend of the data
- Change algorithm or features

Over-fitting

- Model fits too well does not “generalize”
- Captures the noise or outliers of the data
- Change algorithm or filter outliers
Why Not Decision Trees

- Slightly prone to over-fitting
  - But we’ll fix this with ensembles
- Splitting prefers decision boundaries that are parallel to feature axes
Splits Parallel to Axis

Ideal split…
But not Possible!
Will “discover” diagonal edge eventually
Why Not Decision Trees

• Slightly prone to over-fitting
  • But we’ll fix this with ensembles
• Splitting prefers decision boundaries that are parallel to feature axes
  • More data!
• Predictions outside training data can be problematic
Outlier Predictions
Why Not Decision Trees

- Slightly prone to over-fitting
  - But we’ll fix this with ensembles
- Splitting prefers decision boundaries that are parallel to feature axes
  - More data!
- Predictions outside training data can be problematic
  - We can catch this with model competence
- Can be sensitive to small changes in training data
Outlier Predictions
Why Not Decision Trees

- Slightly prone to over-fitting
  - But we’ll fix this with ensembles
- Splitting prefers decision boundaries that are parallel to feature axes
  - More data!
- Predictions outside training data can be problematic
  - We can catch this with model competence
- Can be sensitive to small changes in training data

Questions:
- What other models can we try?
- And how will we know which one works best?
Your turn!

• Build a **Model** for the PDX dataset
  • Use the same dataset from before!
  • This will be a regression
• What is the most important feature?
• What is the predicted price of:
  • SQFT = 1,000, LOT SIZE=0
  • All other fields at default
• Can you disable all fields? What prediction is that?
• How does the prediction change with Missing Strategy: Last versus Proportional