Non-Parametric Models
Review of last class:
Decision Tree Learning

- dealing with the overlearning problem: pruning
- ensemble learning
- boosting
Agenda

• Nearest neighbor models
• Finding nearest neighbors with kd trees
• Locality-sensitive hashing
• Nonparametric regression
Non-Parametric Models

• doesn’t mean that the model lacks parameters
• parameters are not known or fixed in advance
• make no assumptions about probability distributions
• instead, structure determined from the data
Comparison of Models

<table>
<thead>
<tr>
<th>Parametric</th>
<th>Non-Parametric</th>
</tr>
</thead>
<tbody>
<tr>
<td>data summarized by a</td>
<td>data summarized by an unknown (or non-fixed)</td>
</tr>
<tr>
<td>fixed set of parameters</td>
<td>set of parameters</td>
</tr>
<tr>
<td>once learned, the</td>
<td>must keep original data</td>
</tr>
<tr>
<td>original data can be</td>
<td>to make predictions or</td>
</tr>
<tr>
<td>discarded</td>
<td>to update model</td>
</tr>
<tr>
<td>good when data set is</td>
<td>may be slower, but</td>
</tr>
<tr>
<td>relatively small – avoids</td>
<td>generally more accurate</td>
</tr>
<tr>
<td>overfitting</td>
<td></td>
</tr>
<tr>
<td>best when correct</td>
<td></td>
</tr>
<tr>
<td>parameters are chosen!</td>
<td></td>
</tr>
</tbody>
</table>
Instance-Based Learning
Decision Trees

- examples (training set) described by:
  - input: the values of attributes
  - output: the classification (yes/no)
- can represent any Boolean function
Another NPM approach: Nearest neighbor (k-NN) models

- given query $x_q$
- answer query by finding the $k$ examples nearest to $x_q$
- classification:
  - take plurality vote (majority for binary classification) of neighbors
- regression
  - take mean or median of neighbor values
Example: Earthquake or Bomb?
Modeling the data with k-NN

$k = 1$

$k = 5$
Measuring “nearest”

• Minkowski distance calculated over each attribute (or dimension) $i$

$$L^p(x_j, x_q) = \left( \sum_i |x_{j,i} - x_{q,i}|^p \right)^{1/p}$$

• $p = 2$: Euclidean distance – typically used if dimensions measure similar properties (e.g., width, height, depth)
• $p = 1$: Manhattan distance – if dimensions measure dissimilar properties (e.g., age, weight, gender)
Recall a problem we faced before

- shape of the data looks very different depending on the scale
  - e.g., height vs. weight, with height in mm or km
- similarly, with k-NN, if we change the scale, we’ll end up with different neighbors
Simple solution

- simple solution is to normalize:

\[ x'_{j,i} = \frac{(x_{j,i} - \mu_i)}{\sigma_i} \]
Example: Density estimation

128-point sample

MoG representation

smallest circles enclosing 10 neighbours
Density Estimation using k-NN

- # of neighbours impacts quality of estimation

ground truth

k=3

k=10

k=40
Curse of dimensionality

- we want to find $k = 10$ nearest neighbors among $N=1,000,000$ points of an $n$-dimensional space
- sounds easy, right?
- volume of neighborhood is $k/N$
- average side length $l$ of neighborhood is $(k/N)^{1/n}$

<table>
<thead>
<tr>
<th>$n$</th>
<th>$l$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.00001</td>
</tr>
<tr>
<td>2</td>
<td>.003</td>
</tr>
<tr>
<td>3</td>
<td>.002</td>
</tr>
<tr>
<td>10</td>
<td>.3</td>
</tr>
<tr>
<td>20</td>
<td>.56</td>
</tr>
</tbody>
</table>
k-dimensional (kd) trees

- balanced binary tree with arbitrary # of dimensions
- data structure that allows efficient lookup of nearest neighbors (when # of examples >> k)
- recursively divides data into left and right branches based on value of dimension $i$
k-dimensional (kd) trees

- query value might be on left half of divide but have some of $k$ nearest neighbors on right half
- decide whether to inspect the right half based on distance of best match found from dividing hyperplane
Locality-Sensitive Hashing (LSH)

- uses a combination of $n$ random projections, built from subsets of the bit-string representation of each value
- value of each of the $n$ projections stored in the associated hash bucket
Locality-Sensitive Hashing (LSH)

- on search, the set of points from all hash buckets corresponding to the query are combined together
- then measure distance from query value to each of the returned values
- real-world example:
  - data set of 13 million samples of 512 dimensions
  - LSH only needs to examine a few thousand images
  - 1000-fold improvement over kd trees!
Nonparametric Regression Models

• Let’s see how different NPM strategies fare on a regression problem
Piecewise linear regression
3-NN Average
Linear regression through 3-NN
Local weighting of data with kernel

quadratic kernel with $k = 10$:

$$w = \max(0, 1 - \left(\frac{2|x|}{k}\right)^2)$$
Locally weighted quadratic kernel $k=10$
Comparison

connect the dots

3-NN linear regression

3-NN average

locally weighted regression (quadratic kernel width k=10)
Next class

• Statistical learning methods, Ch. 20