CS 106A, Lecture 26
A Gentle Intro. to Machine Learning

suggested reading:
none!
Plan for today

• What is machine learning?
• Why is it useful?
• Supervised Learning
  – Demo: k-Nearest-Neighbor
• What else can machine learning do?
Animal Classification Ex.

• We have a picture and we want to know if it’s a cat or not.

- True
- False
- True
- False
private void isCat(GImage animal) {
    int[][][] pixels = animal.getPixelArray();
    if (containsTwoEyes(pixels)) {
        if (hasWhiskers(pixels)) {
            if (hasPointyEars(pixels)) {
                return true;
            }
        }
    }
    return false;
}
Some tricky cases
Pros/Cons

• Pros
  – Matches our human intuition about what a cat is
  – Easy to understand the code

• Cons
  – Requires us to explicitly enumerate every feature that’s important, and know how important it is
  – Need to write code to detect eyes, and whiskers, and the pointiness of ears
  – Will never improve... cannot learn from its mistakes
What is Machine Learning?

• “The field of study that gives computers the ability to learn without being explicitly programmed” - Arthur Samuel, 1959

• How can a computer do this?

Data.
• Here’s a sketch of how we can approach this problem in the machine learning paradigm:

• Input to the algorithm: MANY cats and MANY not-cats
• For each image provided to the algorithm:
  – Predict whether image is cat or not-cat
    • If correct, great, proceed to the next image
    • If incorrect, **update** the algorithm to do better next time

* * *

**cat**  
**cat**  
* * *  
**not cat**
Pros/Cons

• Pros
  – Doesn’t require us to explicitly tell the algorithm what’s important to distinguish between cats and not-cats
  – Not specific to the cat problem: we could show it images of anything and train it to learn the difference
  – Gets better the more data we give it

• Cons
  – Sometimes hard to know why the algorithm makes the predictions it does
  – Requires us to specify an update mechanism: how is the algorithm supposed to improve itself?
  – Might require a lot of data to perform well
Where is this useful?

• Skin cancer classification
  – Is a given lesion **benign** or **malignant**?

• A machine learning algorithm has been shown to perform **as well** as dermatologists.

• There are other tasks where machine learning algorithms perform **better** than skilled humans.

Digit Classification

• Task:
  – Given a picture of a handwritten digit from 0-9, predict which integer it is
Digit Classification

• Task:
  – Given a picture of a handwritten digit from 0-9, predict which integer it is

\[ 3 \rightarrow 3 \quad 4 \rightarrow 4 \quad 2 \rightarrow 2 \quad 9 \rightarrow 9 \quad \text{“label”} \]
The algorithm gets **training data** that it can use to make predictions.

We use **test data** to evaluate how well it performs.
k-Nearest Neighbors

• Idea: when given a test image, look through all the training images to find the “closest” image.

• Under the assumption that “close” images share the same label, return the label of that closest training image.

  – What does it mean for an image to be “close” to another image?
**k-Nearest Neighbors**

- Idea: when given a **test** image, look through all the **training** images to find the **k** “closest” images.

- Under the assumption that “close” images share the same label, return the **most common** label of the **k closest** images.
k-Nearest Neighbors

Given a test image $T$ and a set of training images $S$:

$C = k$ closest images to $T$

For each training image $I$ in $S$:

$$\text{distance} = \text{distance}(I, T)$$

if $\text{distance} < \text{distance}$ between $I$ and the images in $C$:

add $I$ to $C$

return most common label in $C$
k-Nearest Neighbors

TRAIN

\[3 \ 1 \ 4 \ 0 \ 4 \ 4 \ 1 \ 2 \ 9 \ 5 \ 3 \ 4\]

TEST

K=3

4
k-Nearest Neighbors

TRAIN

3 1 4 0 4 4
1 2 9 5 3 4

TEST

K=3

4
k-Nearest Neighbors

Given a test image $T$ and a set of training images $S$:

$C = k$ closest images to $T$

For each training image $I$ in $S$:

$\text{distance} = \text{distance}(I, T)$

if $\text{distance} < \text{distance}$ between $I$ and the images in $C$: add $I$ to $C$

return most common label in $C$
nothing to see here, carry on 😊
What else can ML do?

• Supervised Learning
  – Classification, like the cat or skin cancer or digits examples
  – Regression
Regression Example

- **House price prediction**

<table>
<thead>
<tr>
<th>Square Footage</th>
<th>House Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,000</td>
<td>150,000</td>
</tr>
<tr>
<td>1,256</td>
<td>175,000</td>
</tr>
<tr>
<td>5,897</td>
<td>2,000,000</td>
</tr>
<tr>
<td>4,300</td>
<td>1,300,000</td>
</tr>
<tr>
<td>2,400</td>
<td>750,000</td>
</tr>
<tr>
<td>2,600</td>
<td>690,000</td>
</tr>
<tr>
<td>3,000</td>
<td>1,000,000</td>
</tr>
</tbody>
</table>

**Training Data**

**Test Time**

“My house has 1,800 square feet. How much should I expect it to sell for?”
What else can ML do?

• Supervised Learning
  – Classification
  – Regression

• Unsupervised Learning
  – Finding structure in unlabeled data
Netflix Prize Example

• Netflix Prize
What else can ML do?

• Supervised Learning
  – Classification
  – Regression

• Unsupervised Learning

• Reinforcement Learning
What remains really hard?

• Human-like dialog
• Common-sense reasoning about the world
• Strong generalization
• Learning to learn
• ... and much else 😊