CSE 151 Intro. to AI: A Statistical Approach

Instructor: Kamalika Chaudhuri
CSE 151 Machine Learning

Instructor: Kamalika Chaudhuri
Logistics

**Instructor:** Kamalika Chaudhuri  
**Office Hours:** MW 3:30-4:30pm, 4110 EBU3B  
**Email:** kamalika@cs.ucsd.edu

**TA:** Akshay Balasubramani  
**Office Hours:** TBD  
**Email:** abalsubr@cs.ucsd.edu

**TA:** Yuncong Chen  
**Office Hours:** TBD  
**Email:** yuncong@cs.ucsd.edu
What is Machine Learning?

How to use data to learn to make better predictions

Example 1: Recommendation Systems

Suggestions to Watch Instantly

Wallace & Gromit: Loaf and Death
Because you enjoyed:
The Iron Giant
Howl's Moving Castle
Coraline

Iron Man: Armored Adventures
Because you enjoyed:
The Iron Giant
How to Train Your Dragon
Kung Fu Panda

Traffik
Because you enjoyed:
This Is Spinal Tap
The Third Man
Rashomon
What is Machine Learning?

How to use data to learn to make better predictions

Example 2: Spam Detection
What is Machine Learning?

How to use data **to learn** to make better predictions

Example 3: Link Prediction
What is Machine Learning?

How to use data to learn to make better predictions
Algorithm behavior changes based on data

This class: some basic machine learning methods
Two Types of Machine Learning

**Supervised Learning**
Given examples of data and their labels, predict labels of new (unseen) data

**Unsupervised Learning**
Given data, build a model or cluster

There are other types, but we won’t get to it in this class
Supervised Learning

Classification:
Given labeled data:

\[(x_i, \ y_i)\]  \(i=1,...,n\)

where \(y\) is **discrete**, find a rule to predict \(y\) values for **unseen** \(x\)
Typical Classification Algorithm

Set of input examples \((x_i, y_i)\)

\[ \text{Classification Algorithm} \]

\[ \text{Prediction Rule} \]

New example \(x\)

Label \(y\)
Typical Classification Algorithm

Set of input examples \((x_i, y_i)\)

Classification Algorithm

Prediction Rule

Test Data

New example \(x\)

Label \(y\)

Training Data

Training and test data must be separate!
Typical Classification Algorithm

Set of input examples \((x_i, y_i)\)

Training Data

Classification Algorithm

Prediction Rule

Test Data

New example \(x\)

Label \(y\)

Performance Measure:
Accuracy (or fraction of correct answers) on test data
Supervised Learning

**Classification:** Given labeled data \((x_i, y_i)\) where \(y\) is discrete, predict \(y\) values for unseen \(x\)

**Example 1:** Predict if a **new** patient has flu or not, based on **existing** patient data

What is \(x\) and \(y\)?
**Supervised Learning**

**Classification:** Given labeled data \((x_i, y_i)\) where \(y\) is **discrete**, predict \(y\) values for unseen \(x\)

**Example 1:** Predict if a patient has flu or not

<table>
<thead>
<tr>
<th>Fever</th>
<th>Cold</th>
<th>Temperature</th>
<th>Flu?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>No</td>
<td>99F</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Features: Properties of patient
Label: Flu/No flu

A **binary** (two-label) classification problem
Supervised Learning

**Classification:** Given labeled data \((x_i, y_i)\) where \(y\) is *discrete*, predict \(y\) values for unseen \(x\)

**Example 2:** Which digit in the image?

```
0 1 2 3 4
5 6 7 8 9
```

Label: 0, 1, .., 9
What are the features?

A *multiclass* classification problem
Supervised Learning

**Classification**: Given labeled data \((x_i, y_i)\) where \(y\) is discrete, predict \(y\) values for unseen \(x\)

**Example 2**: Which digit in the image?

Label: 0, 1, ..., 9

What are the features?
Option: vector of pixel colors

Image

\[
\begin{array}{cccccc}
0 & 1 & 2 & 3 & 4 \\
5 & 6 & 7 & 8 & 9 \\
\end{array}
\]

\(x\) (0 for white, 1 for black)
Supervised Learning

**Classification:** Given labeled data \((x_i, y_i)\) where \(y\) is **discrete**, predict \(y\) values for unseen \(x\)

**Example 2:** Which digit in the image?

![Digits](image)

Label: 0,1,...,9

What are the features?

Option: vector of pixel colors

There are other options too

**Lesson:** Choosing features is non-trivial in real applications
Supervised Learning

**Classification:** Given labeled data \((x_i, y_i)\) where \(y\) is **discrete**, predict \(y\) values for unseen \(x\)

**Example 3:** Spam or not?

<table>
<thead>
<tr>
<th>Email 1</th>
<th>Email 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>From: Canadian Pharmacy</td>
<td>From: Yuncong Chen</td>
</tr>
<tr>
<td>Subject: Offer ends now!</td>
<td>Subject: TA meeting</td>
</tr>
<tr>
<td><em>Pharmacy</em></td>
<td><em>Spam?</em></td>
</tr>
<tr>
<td><em>offer</em></td>
<td></td>
</tr>
<tr>
<td><em>meeting</em></td>
<td></td>
</tr>
<tr>
<td><em>TA</em></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td></td>
</tr>
</tbody>
</table>

Label: 0 (not spam), 1 (spam)  Features: Words in the email
Supervised Learning

Regression:
Given data:

\[(x_i, \ y_i) \quad \text{i=1,...,n}\]

\(x_i\) independent variable
\(y_i\) dependent variable

where \(y\) is **continuous**, design a rule to predict \(y\) values for unseen \(x\)
Supervised Learning

Regression: Given data \((x_i, y_i)\)
where \(y\) is continuous, predict \(y\) values for unseen \(x\)

Example 1: Predict house price from properties of house

<table>
<thead>
<tr>
<th>Bedrooms</th>
<th>Bathrooms</th>
<th>Area</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>2</td>
<td>2000</td>
<td>600K</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1200</td>
<td>400K</td>
</tr>
</tbody>
</table>

Independent Variable: Property of house
Dependent variable: price
Two Types of Machine Learning

Supervised Learning
Given examples of data and their labels, predict labels of new (unseen) data
Examples: Classification, Regression

Unsupervised Learning
Given data, build a model

There are other types, but we won’t get to it in this class
Unsupervised Learning

Clustering
Given a set of input objects, group them to clusters by similarity

Example 1: Cluster videos by people in them
Unsupervised Learning

Clustering
Given a set of input objects, group them to clusters by similarity

Example 2: Cluster documents by topic

Physics
  - Gravity
  - Laws of Motion
  - Electricity

Math
  - Geometry
  - Algebra

Features: Words in the document
Unsupervised Learning

Dimensionality Reduction
Given high dimensional data, find a good low dimensional representation

Example 1: Images

Number of pixels = 768, so 768-dimensional object
Can we find a lower dimensional representation?
Two Types of Machine Learning

**Supervised Learning**
Given examples of data and their labels, predict labels of new (unseen) data
Examples: Classification, Regression

**Unsupervised Learning**
Given data, build a model
Examples: Clustering, Dimension Reduction, learning HMMs

There are other types, but we won’t get to it in this class
Logistics

**Instructor:** Kamalika Chaudhuri  
**Email:** kamalika@cs.ucsd.edu

**Lecture:** MW 5-6:20pm, WLH 2204  
**Sections (optional):** W10-10:50am, PETER 104  
No section Wed Apr 4

**Website:** [http://cseweb.ucsd.edu/classes/sp12/cse151-a/](http://cseweb.ucsd.edu/classes/sp12/cse151-a/)
Logistics

Instructor: Kamalika Chaudhuri
Office Hours: MW 3:30-4:30pm, 4110 EBU3B
Email: kamalika@cs.ucsd.edu

TA: Akshay Balasubramani
Office Hours: TBD
Email: abalsubr@cs.ucsd.edu

TA: Yuncong Chen
Office Hours: TBD
Email: yuncong@cs.ucsd.edu
Administrivia

Prerequisites:
Some probability, linear algebra and calculus
CSE 150 helps, but absolutely essential

Textbooks:
No textbook for this class

Syllabus:
Classification -- k-NN, Perceptron, Boosting, etc.
Learning Theory
Unsupervised learning -- k-means, GMM, PCA, etc.
Assessment

Homeworks (4): 25%
Midterms (2): 30%
Final: 40%
Class Participation: 5%

We will be using GradeSource
Homework Policy

Homeworks are due in class at the beginning of lecture

No late homeworks will be accepted

Homework with the lowest grade will be dropped
Programming Assignment Policy

Homeworks will be a mix of programming + pen and paper assignments

You can use any language and any libraries for your programming assignments

If you use external libraries, it is your responsibility to make sure they give you correct answers

Submit a printout of your code with your homework
Collaboration Policy

Homeworks should be done in **groups of two**

You are not permitted to collaborate with anyone outside your homework group

Email me the name of your homework partner by April 11

If you need a homework partner, send me email by April 11, and I will pair you up
Exam Policy

Midterms and final will be closed book, closed notes
You can bring a “crib-sheet” with you
The first midterm is on April 23
Class Participation

Class participation is very important for this class.

In some lectures, we will have mini problem-sessions.

Your class participation grade depends on how you do in those sessions.

Bring blank sheets of paper to class for problem sessions.

Ask lots of questions, if you don’t understand something!
Questions

Message board for this class on Piazza

Please post your questions on the message board!
Academic Honesty

**Remember:**
Academic dishonesty is taken very seriously in this class

More details on the class website
Typical Classification Algorithm

Training Data

Set of input examples \((x_i, y_i)\)

Classification Algorithm

Prediction Rule

Test Data

New example \(x\)

Label \(y\)

Training and test data must be separate!
Generative Classification

**Goal:** Classify red from blue

**Generative:**
Model each class probabilistically
Learn the parameters of each class from data

For a test example $x$, find $P(\text{class 1}|x)$ and $P(\text{class 2}|x)$
Report 1 if $P(\text{class 1}|x) > P(\text{class 2}|x)$, 2 otherwise
Discriminative Classification

**Goal:** Classify *red* from *blue*

![Graph showing discriminative classification with a linear separator between red and blue points.]

**Discriminative:**
No need to model each class probabilistically
Find a suitable separator (say a linear separator) that mostly separates red from blue

Advantages and Disadvantages?