CSE 151 Intro. to AI: A Statistical Approach

Instructor: Kamalika Chaudhuri
Course Staff

Instructor: Kamalika Chaudhuri  (kamalika@cs)
Office Hours:  F1-2pm, CSE 4110

TAs:

    Chicheng Zhang     Joseph Geumlek
    Shuang Song       Songbai Yan
    Yizhen Wang

Tutor:   Alvin See

Website:

    http://cseweb.ucsd.edu/classes/wi17/cse151-a/
What is Machine Learning?

How to use data to learn to make better predictions

Example 1: Recommendation Systems
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:

$$(x_i, \quad y_i) \quad i=1,\ldots,n$$

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

Set of input examples \((x_i, y_i)\)  \rightarrow Classification Algorithm  \rightarrow Prediction Rule

New example \(x\)  \rightarrow Label \(y\)
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!
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 \textit{discrete}, predict \(y\) values for unseen \(x\)

Example 1: Predict if a \textit{new} patient has flu or not, based on \textit{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>
<tr>
<td>1</td>
<td>0</td>
<td>99</td>
<td>+</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

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

\[
\begin{array}{cccccc}
\mathbf{0} & 0 & 1 & 0 & 1 & 0 & 1 \\
\ldots \\
\end{array}
\]

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?

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>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pharmacy</th>
<th>offer</th>
<th>meeting</th>
<th>TA</th>
<th>Spam?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Email 1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Email 2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

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

Regression:

Given data:

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

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

0 1 2 3 4
5 6 7 8 9

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: MWF 11-11:50 (Sec A), 12-12:50 (Sec B)
Sections (optional): M6-6:50pm (Sec A), 7-7:50pm (Sec B), Center 214

Website: http://cseweb.ucsd.edu/classes/wi17/cse151-a/
Textbooks:
No textbook for this class

Syllabus:
Classification -- k-NN, Perceptron, Boosting, etc.
Linear Least Squares Regression
Unsupervised learning -- k-means, hierarchical clustering
Prerequisites

Probability: Events, random variables, expectations, joint, conditional and marginal distributions, independence

Linear Algebra: Vector spaces, subspace, matrix inversion, matrix multiplication, linear independence, rank, determinant, bases, orthonormality, solving systems of linear equations

Calculus: Minima, maxima of functions, derivatives, integrals

Programming: Write programs in a language of your choice. No hand-holding provided
Prerequisites

Calibration homework HW0 will go out on Fri!
Due in lecture on Jan 20

HW0 covers most (but not all) of the material you need to know as a pre-requisite
Assessment

Homeworks (7): 30%
Midterm: 30%
Final: 35%
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

Homeworks will be graded based on correctness and clarity
Homeworks

Lectures will cover background conceptual material you need to do your assignments

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**

Email us a copy of your code with your HW (instructions later)
Collaboration Policy

Homeworks should be done in groups of one, two or three.

All members in a group should be enrolled in the same section.

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

Email me the name of your homework partner by Jan 20.

If you need a homework partner, please post on Piazza.
Midterms

The midterm will be held in class on Wed Feb 8

To get any credit, you should take the midterm for the section you are enrolled in
Questions

Message board for this class on Piazza

Please post your questions on the message board!
Typical Classification Algorithm

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

Training Data

Classification Algorithm

Prediction Rule

New example \(x\)

Test Data

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
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

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?
Generative vs. Discriminative

This class we will mostly cover discriminative models