20180111 COGS 118a Lecture Notes

Cabinet COGS118a Lecture Notes

Iris dataset

Perform classification on different Iris based on features.

2d feature set

Sepal Width vs. Sepal Length

Fashion MNIST

Dataset for classifying 10 different types of clothes

t-SNE is used to help visualize clusters for high dimensional space problems (500 to 2).

Amazon

Machine Learning uses your purchase history to help predict what you're likely to buy in the future.

This means Amazon could potentially ship products to nearby distribution centers before you even decide to buy.

Pipeline

1. Find the data source
2. Crawl the data
3. Perform data cleansing
4. Data processing and visualization
5. Training your machine learning algorithm

Example

Amazon purchase review

- Rating
- Helpfulness

Find the data source

API, not always available

Scrape the site (Srapy)

Data Cleansing

- Incomplete
- noisy
- inconsistent
- intentional (disguised missing data)

Process the data

Train your classifier

Draw a line through your data (decision boundary)

What is a pattern

- repetitive
- common features
- subjective
- Explicit and implicit descriptions

“Everything is a pattern”

No pattern is also a pattern

Key concepts in Machine Learning

- Representation
- Evaluation
- Optimization
- Data
- Computing power

Notation

$S = \{x_i, i = 1..n\}$: A set S with n samples. i goes from 1 to n.

> Order in a set does **not** matter.

$\mathbf{x}_i = (x_{i1}, \ldots, x_{im})$: A row vector of m elements

Sometimes we'll use brackets: $\mathbf{x}_i =< x_{i1}, \ldots, x_{im} >$

Training

$S_{training} = \{(\mathbf{x}_i, y_i), i = 1..n\}$

In supervised setting during training, y_i (the solution) to each sample x_i is provided.

Testing

$S_{testing} = \{(\mathbf{x}_i), i = 1..u\}$, what is y_i?

Supervised Learning

Classification

Decision Boundary

Classify based on whether a data point lies above or below a line in the graph.

Regression

Draw a line through your data points. This is similiar to the decision boundary, hence, we dont' differentiate between regression and classification.

Unsupervised Learning

No Labels

Clustering

Separate data into clusters

Clustering results are **not** unique

Representing a raw image

$n \times m$ matrix gets turned into a vector.

For now…
Every single input is a vector

Mathematical Representation for features

Binary encoding: male = 0 or female = 1

One-hot encoding: San Diego = 100, Irving = 010, LA = 001

We don't assign arbitrary integers to categories, i.e. San Diego = 0, Irvine = 1, LA = 2, etc.. because this causes mathematical problems down the road. We want to abstract the mathematical representation of the feature from the definition of the feature.

The cost with doing this is that now there is a column per feature. The feature dimension becomes much larger.

Input matrix

$$S = \{x_i, i = 1..n\} \quad x_i = (x_{i1}, \ldots, x_{im})$$

as a column vector…

$$X = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

or as a row vector…

$$X = (x_1 x_2 x_3)$$