CS 340: Machine Learning
Lecture 2: Introduction to Supervised Learning

AD

January 2011

Supervised Learning

- Given a training set of N input-output pairs $\{\mathbf{x}^i, y^i\} \in \mathcal{X} \times \mathcal{Y}$, "learn" a function $f : \mathcal{X} \to \mathcal{Y}$ to predict the output $\widehat{y} = f(\mathbf{x})$ associated to a new input $\mathbf{x}$.
  - Each input $\mathbf{x}^i$ is a p-dimensional feature vector (covariates, explanatory variables).
  - Each output y^i is a target variable (response).
- Classification corresponds to $\mathcal{Y} = \{1, ..., K\}$.
- Regression corresponds to $\mathcal{Y} = \mathbb{R}^d$.
- **Aim**: produce the correct output given a new input.

Practical examples of classification

- Email spam filtering (feature vector = "bag of words").
- Webpage classification ("bag of words", URL etc).
- Detecting credit card fraud (#transactions, average transactions, locations).
- Credit scoring (income, saving, degree, age...)
- Handwritten digit recognition.

Handwritten digit recognition

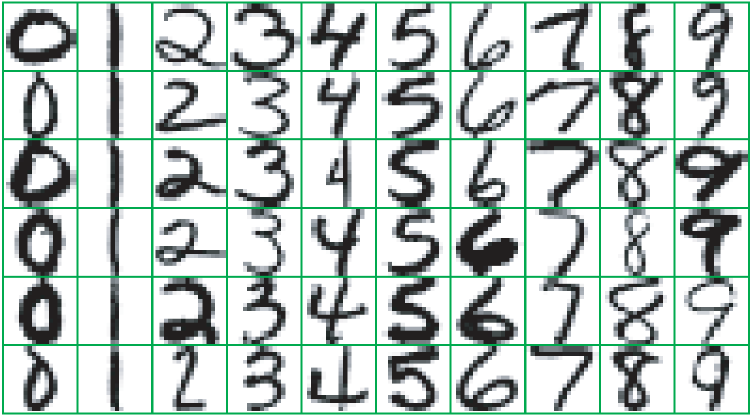


Figure: Examples of handwritten digits from US postal employes

- In this case, $\mathcal{X} = \{0, 1\}^{16 \times 16}$ and $\mathcal{Y} = \{0, 1, ..., 9\}$.

Recognizing Tufas

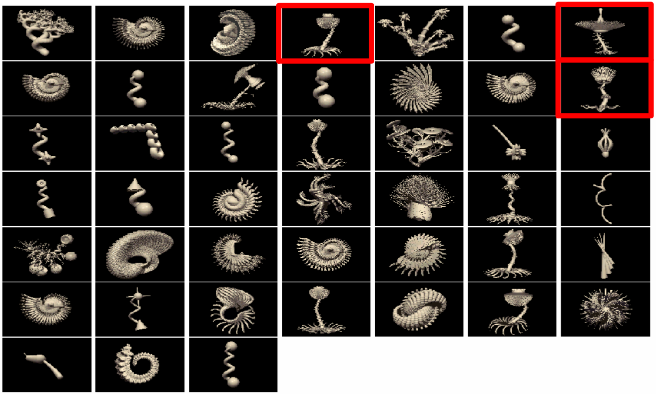
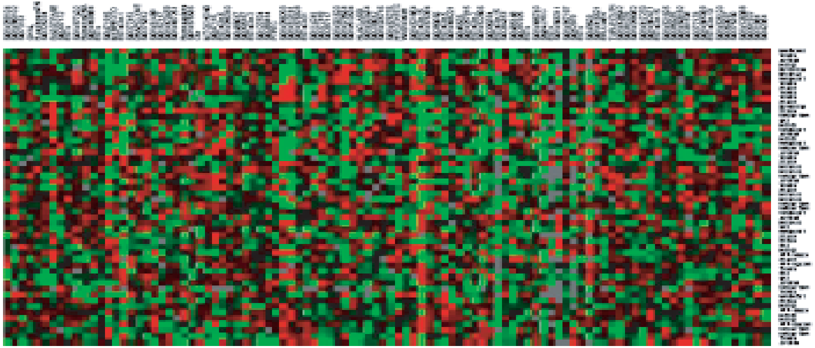


Figure: Can you pick out the tufas?

- In this case, $\mathcal{X} = \{0, ..., 255\}^{128 \times 128}$ and $\mathcal{Y} = \{0, 1\}$.

Classifying Gene Microarrays for Cancer Diagnosis

- Training data: $\mathbf{x}^i$ gene expression data on p genes and $y^i \in \{0, 1\}$ (cancer/no cancer).



- In this case, we have $\dim\left(\mathbf{x}^i\right) >> N$.

Learning is plagued with problems

- Is there any information present in the data? (e.g. are you monitoring the right genes?)
- Is there enough information in the data? (e.g. are you monitoring only a part of the genes?)
- Is there too much/irrelevant information in the data? (e.g. are you monitoring all the genes? FDR).
    - **True example 1**: There is a close relationship between the salaries of Presbyterian ministers in Massachusetts and the price of rum in Havana.
    - **True example 2**: Connect neuroimaging data to measures of behavior found in social and cognitive neuroscience. Some researchers do one correlation analysis against all the voxels in the brain (~160,000+) to find those that are related to their measure of behavior.
- Training data are noisy and/or mislabelled (e.g. measurement errors/diagnosis error).

Supervised learning as function fitting

- We are given some training data

$$\mathcal{D} = \left\{ \left(\mathbf{x}^i, y^i \right) \right\}_{i=1}^{N}$$

- Consider a restricted set of mappings/parametric functions f in *hypothesis class* $\mathcal{H}$

$$f \in \mathcal{H} : \mathcal{X} \times \quad \Theta \quad \longrightarrow \mathcal{Y},$$

  we will predict using

$$\widehat{y}\left(\mathbf{x}\right) = f\left(\mathbf{x}; \theta\right)$$

  where $\theta \in \Theta$.

- **Learning**: Given $\mathcal{H}$, learn parameters θ given $\mathcal{D}$ so that predictions on non-labeled inputs (i.e. test set, real-world data) are as accurate as possible.

Definitions

- **Training error**:

$$\text{Err\_Train} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(\widehat{y}\left(\mathbf{x}^{i}\right) \neq y^{i}\right).$$

- **Test error**:

$$\text{Err\_Test} = \frac{1}{N\_\text{test}} \sum_{i=1}^{N\_\text{test}} \mathbb{I}\left(\widehat{y}\left(\mathbf{x}_{\text{test}}^{i}\right) \neq y_{\text{test}}^{i}\right).$$

- Test error cannot be computed in real-world applications where $\left\{y_{\text{test}}^{i}\right\}$ is not available.

Binary classification: Credit card scoring

- Say you have training data of the following form

Income	Savings	Risk
100	50	Hi
100	100	Lo
50	75	Hi
500	93	Lo

- Test data are of the form

Income	Savings	Risk
98	49	?
100	102	?
400	20	?

- In this case $\mathbf{x} = (x_1, x_2) \in \left(\mathcal{X} = \mathbb{R}^2\right)$ and $\mathcal{Y} = \{\text{high,low}\}$.

Example function

- $f(\mathbf{x};\theta) = f((\text{income,savings});(\theta_1,\theta_2)) =$ IF $(x_1 = \text{income}) > \theta_1$ AND $(x_2 = \text{savings}) > \theta_2$ THEN low-risk ELSE high-risk.