Stat 406: Algorithms for classification and prediction

Lecture 1: Introduction

Kevin Murphy

Wed 4 January, 2005\footnote{Slides last updated on January 4, 2006}
OUTLINE

• Administrivia
• Machine learning: some basic definitions.
• Simple examples of regression.
• Real-world applications of regression.
• Simple examples of classification.
• Real-world applications of classification.
**Administrivia**

- Web page
  

- Please fill out the sign-up sheet.

- Labs Wed 4-5.

- The TA is Aline Tabet.

- My office hours are Fri 2-3pm LSK 308d.

- Aline’s office hours are TBA (see web).
Grading

- There will be weekly homework assignments worth 20%. Out on Mondays, return on Mondays (in class).
- The homeworks will often involve programming; you may want to do this part during the lab.
- The midterm will be in late February and is worth 40%.
- The final will be in April and is worth 40%.
Pre-requisites

- Math: multivariate calculus, linear algebra, probability theory.
- Stats: stats 306 or CS 340 or equivalent.
- CS: some experience with programming (e.g. in R) is required.
MATLAB

• There will be weekly programming assignments (as part of the lab).
• We will use matlab for programming.
• Matlab is very similar to R, but is somewhat faster and easier to learn. Matlab is widely used in the machine learning and Bayesian statistics community.
• Unfortunately matlab is not free (unlike R). You can buy a copy from the bookstore for $150, or you can use the copy installed in the lab machines.
• You will learn how to use matlab during the first few labs.
• There is no official textbook. I will hand out various notes in class, including some chapters from the following unfinished/ unpublished books
  – *Probabilistic graphical models*, Michael Jordan, 2006
The following (already published) books are also recommended
  – *Elements of statistical learning*, Hastie, Friedman and Tibshirani, 2001. (Available from the bookstore)
• Since this is a new course, the syllabus is likely to change during the course of the semester.

• See the web page for details.

• You will get a good feeling for the class during today’s lecture.
● Administrivia √
● Machine learning: some basic definitions.
● Simple examples of regression.
● Real-world applications of regression.
● Simple examples of classification.
● Real-world applications of classification.
LEARNING TO PREDICT

• This class is about supervised approaches to machine learning.
• Given a training set of $n = N_D$ input-output pairs $D = (\vec{x}_i, \vec{y}_i)_{i=1}^{N_D}$, we attempt to construct a function $f$ which will accurately predict $f(\vec{x}_*)$ on future, test examples $\vec{x}_*$.
• Each input $\vec{x}_i$ is a vector of $p = N_X$ features or covariates. Each output $\vec{y}_i$ is a target variable. The training data is stored in an $N_D \times N_X$ design matrix $X = [\vec{x}_i^T]$. The training outputs are stored in a $N_D \times N_Y$ matrix $Y = [\vec{y}_i^T]$. 

![Diagram of design matrix and output matrix]
Classification vs regression

- If $\vec{y} \in \mathbb{R}^{N_Y}$ is a continuous-valued output, this is called regression. Often we will assume $N_Y = 1$, i.e., scalar output.
- If $y \in \{1, \ldots, N_Y\}$ is a discrete label, this is called classification or pattern recognition. The labels can be ordered (e.g., low, medium, high) or unordered (e.g., male, female). $N_Y$ is the number of classes. If $N_Y = 2$, this is called binary (dichotomous) classification.
Notation

- We will denote discrete ranges \( \{1, \ldots, N\} \) by \( 1 : N \).
- We will often encode categorical variables as binary vectors. eg if \( y \in \{1, \ldots, 3\} \), we will use \( \vec{y} \in \mathbb{0,1}^3 \), where \( y = 1 \) maps to \( \vec{y} = (1, 0, 0) \), \( y = 2 \) maps to \( \vec{y} = (0, 1, 0) \), and \( y = 3 \) maps to \( \vec{y} = (0, 0, 1) \).
- In general, \( Y = j \) turns bit \( j \) of \( \vec{Y} \) on, with the constraint \( \sum_{j=1}^N Y_j = 1 \).
Short/fat vs tall/skinny data

- In traditional applications, the design matrix is tall and skinny \((n \gg p)\), i.e., there are many more training examples than inputs.
- In more recent applications (e.g., bio-informatics or text analysis), the design matrix is short and fat \((n \ll p)\), so we will need to perform feature selection and/or dimensionality reduction.
Generalization performance

We care about performance on examples that are different from the training examples (so we can’t just look up the answer).
• The *no free lunch theorem* says (roughly) that there is no single method that is better at predicting across all possible data sets than any other method.

• Different learning algorithms implicitly make different assumptions about the nature of the data, and if they work well, it is because the assumptions are reasonable in a particular domain.
Supervised vs unsupervised learning

• In supervised learning, we are given \((\vec{x}_i, \vec{y}_i)\) pairs and try to learn how to predict \(\vec{y}_*\) given \(\vec{x}_*\).

• In unsupervised learning, we are just given \(\vec{x}_i\) vectors.

• The goal in unsupervised learning is to learn a model that “explains” the data well. There are two main kinds:
  – Dimensionality reduction (eg PCA)
  – Clustering (eg K-means)
OUTLINE

• Administrivia ✓
• Machine learning: some basic definitions. ✓
• Simple examples of regression.
• Real-world applications of regression.
• Simple examples of classification.
• Real-world applications of classification.
The output density is a 1D Gaussian (Normal) conditional on $x$:

$$p(y|\vec{x}) = \mathcal{N}(y; \vec{\beta}^T \vec{x}, \sigma) = \mathcal{N}(y; \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p, \sigma)$$

$$\mathcal{N}(y|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2\sigma^2}(y - \mu)^T(y - \mu)\right)$$

For example, $y = ax_1 + b$ is represented as $\vec{x} = (1, x_1)$ and $\vec{\beta} = (b, a)$. 
**Polynomial regression**

If we use linear regression with non-linear basis functions

\[
p(y|x_1) = \mathcal{N}(y|\beta^T [1, x_1, x_1^2, \ldots, x_1^k], \sigma)
\]

we can produce curves like the one below.

Note: In this class, we will often use \( \vec{w} \) instead of \( \vec{\beta} \) to denote the weight vector.
Where to put them? — Segmentation problem.
2D LINEAR REGRESSION

Figure 3.1: Linear least squares fitting with $X \in \mathbb{R}^2$. We seek the linear function of $X$ that minimizes the sum of squared residuals from $Y$. 
PIECEWISE LINEAR 2D REGRESSION

Where to put them? — Segmentation problem.
OUTLINE

• Administrivia ✓
• Machine learning: some basic definitions. ✓
• Simple examples of regression. ✓
• Real-world applications of regression.
• Simple examples of classification.
• Real-world applications of classification.
Real-world applications of regression

• $\vec{x} =$ amount of various chemicals in my factory, $y =$ amount of product produced.
• $\vec{x} =$ properties of a house (eg location, size), $y =$ sales price.
• $\vec{x} =$ joint angles of my robot arm, $\vec{y} =$ location of arm in 3-space.
• $\vec{x} =$ stock prices today, $\vec{y} =$ stock prices tomorrow. (Time series data is not iid, and is beyond the scope of this course.)
• Administrivia ✓
• Machine learning: some basic definitions. ✓
• Simple examples of regression. ✓
• Real-world applications of regression. ✓
• Simple examples of classification.
• Real-world applications of classification.
LINEARLY SEPARABLE 2D DATA

2D inputs $\vec{x}_i \in \mathbb{R}^2$, binary outputs $y \in \{0, 1\}$.
The line is called a decision boundary.
Points to the right are classified as $y = 1$, points to the left as $y = 0$. 
A simple approach to binary classification is logistic regression (briefly studied in 306).

The output density is Bernoulli conditional on $x$:

$$p(y|x) = \pi(x)^y (1 - \pi(x))^{1-y}$$

where $y \in \{0, 1\}$ and

$$\pi(x) = \sigma(\bar{w}^T [1, x_1, x_2])$$

where

$$\sigma(u) = \frac{1}{1 + e^{-u}}$$

is the sigmoid (logistic) function that maps $\mathbb{R}$ to $[0, 1]$. Hence

$$P(Y = 1|\bar{x}) = \frac{1}{1 + e^{-w_0+w_1x_1+w_2x_2}}$$

where $w_0$ is the bias (offset) term corresponding to the dummy column of 1s added to the design matrix.
2D Logistic Regression
Non-linearly separable 2D data

In 306, this is called “checkerboard” data.
In machine learning, this is called the “xor” problem.
The “true” function is $y = x_1 \oplus x_2$.
The decision boundary is non-linear.
Logistic regression with quadratic features

We can separate the classes using

\[ P(Y = 1|x_1, x_2) = \sigma(w^T [1, x_1, x_2, x_1^2, x_2^2, x_1, x_2]) \]
**Outline**

- Administrivia ✓
- Machine learning: some basic definitions. ✓
- Simple examples of regression. ✓
- Real-world applications of regression. ✓
- Simple examples of classification. ✓
- Real-world applications of classification.
Handwritten digit recognition

Multi-class classification.

![Examples of handwritten digits from U.S. postal envelopes.](image)

Figure 1.2: Examples of handwritten digits from U.S. postal envelopes.
Gene microarray expression data

Rows = examples, columns = features (genes).
Short, fat data ($p \gg N$).
Might need to perform feature selection.
Other examples of classification

• Email spam filtering (spam vs not spam)
• Detecting credit card fraud (fraudulent or legitimate)
• Face detection in images (face or background)
• Web page classification (sports vs politics vs entertainment etc)
• Steering an autonomous car across the US (turn left, right, or go straight)