# MATH 567: Mathematical Techniques in Data Science
# Lab 4

Dominique Guillot

Departments of Mathematical Sciences
University of Delaware

February 29, 2017

# Exercise 1

1. Use the command `read.table(file, header = FALSE, sep = " ")` to load the zip codes training and test sets (available on Sakai). Create 4 variables: `train.y`, `train.x`, `test.y`, `test.x`.
   Note: convert `train.y`, `test.y` to "factors" using the `factor` command.
2. Use the `knn` command to predict the labels on the test set using the training set with $k = 5$ neighbors. Compute the prediction error.
3. Install the `caret` and `e1071` packages.
4. Use cross-validation to choose a knn parameter:

```
library(caret)

ctrl <- trainControl(method="repeatedcv", number=10,
                     repeats = 1)

fitKnn = train(train.x, train.y, method="knn",
               trControl = ctrl,
               tuneGrid=expand.grid(.k=1:10),
               metric="Accuracy")
```

5. Compute the prediction error of the "best" knn model.

# Logistic regression

Suppose we work with binary outputs, i.e., $y_i \in \{0, 1\}$.

Linear regression may not be the best model.

- $x^T \beta \in \mathbb{R}$ not in $\{0, 1\}$.
- Linearity may not be appropriate. Does doubling the predictor doubles the probability of $Y = 1$? (e.g. probability of going to the beach vs outdoors temperature).

# Logistic regression

Suppose we work with binary outputs, i.e., $y_i \in \{0, 1\}$.

Linear regression may not be the best model.

- $x^T \beta \in \mathbb{R}$ not in $\{0, 1\}$.
- Linearity may not be appropriate. Does doubling the predictor doubles the probability of $Y = 1$? (e.g. probability of going to the beach vs outdoors temperature).

**Logistic regression:** Different perspective. Instead of modelling the $\{0, 1\}$ output, we model the probability that $Y = 0, 1$.

# Logistic regression

Suppose we work with binary outputs, i.e., $y_i \in \{0, 1\}$.

Linear regression may not be the best model.

- $x^T \beta \in \mathbb{R}$ not in $\{0, 1\}$.
- Linearity may not be appropriate. Does doubling the predictor doubles the probability of $Y = 1$? (e.g. probability of going to the beach vs outdoors temperature).

**Logistic regression:** Different perspective. Instead of modelling the $\{0, 1\}$ output, we model the probability that $Y = 0, 1$.

**Idea:** We model $P(Y = 1 | X = x)$.

- Now: $P(Y = 1 | X = x) \in [0, 1]$ instead of $\{0, 1\}$.
- We want to relate that probability to $x^T \beta$.

Suppose we work with binary outputs, i.e., $y_i \in \{0, 1\}$.

Linear regression may not be the best model.

- $x^T\beta \in \mathbb{R}$ not in $\{0, 1\}$.
- Linearity may not be appropriate. Does doubling the predictor doubles the probability of $Y = 1$? (e.g. probability of going to the beach vs outdoors temperature).

**Logistic regression:** Different perspective. Instead of modelling the $\{0, 1\}$ output, we model the probability that $Y = 0, 1$.

**Idea:** We model $P(Y = 1|X = x)$.

- Now: $P(Y = 1|X = x) \in [0, 1]$ instead of $\{0, 1\}$.
- We want to relate that probability to $x^T\beta$.

We assume

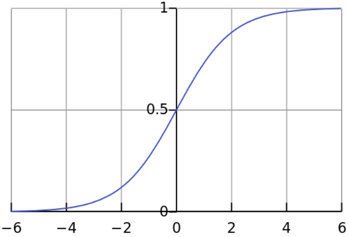
$$\operatorname{logit}(P(Y = 1|X = x)) = \log \frac{P(Y = 1|X = x)}{1 - P(Y = 1|X = x)}$$
$$= \log \frac{P(Y = 1|X = x)}{P(Y = 0|X = x)} = x^T\beta.$$

# Logistic regression (cont.)

Equivalently,

$$P(Y=1|X=x) = \frac{e^{x^T\beta}}{1+e^{x^T\beta}}$$
$$P(Y=0|X=x) = 1 - P(Y=1|X=x) = \frac{1}{1+e^{x^T\beta}}$$

The function $f(x) = e^x/(1+e^x) = 1/(1+e^{-x})$ is called the *logistic function*.




$\log \frac{P(Y=1|X=x)}{P(Y=0|X=x)}$ is the *log-odds* ratio.

- Larger positive values of $x^T\beta \Rightarrow p \approx 1$.
- Larger negative values of $x^T\beta \Rightarrow p \approx 0$.

# Logistic regression (cont.)

In summary, we are assuming:

- $Y|X = x \sim \text{Bernoulli}(p)$.
- $\text{logit}(p) = \text{logit}(E(Y|X = x)) = x^T\beta$.

# Logistic regression (cont.)

In summary, we are assuming:

- $Y|X = x \sim \text{Bernoulli}(p)$.
- $\text{logit}(p) = \text{logit}(E(Y|X = x)) = x^T\beta$.

More generally, one can use a *generalized linear model* (GLM). A GLM consists of:

- A probability distribution for $Y|X = x$ from the exponential family.
- A linear predictor $\eta = x^T\beta$.
- A *link function* $g$ such that $g(E(Y|X = x)) = \eta$.

In logistic regression, we are assuming a model for $Y$. We typically estimate the parameter $\beta$ using maximum likelihood.

# Logistic regression: estimating the parameters

In logistic regression, we are assuming a model for $Y$. We typically estimate the parameter $\beta$ using maximum likelihood.

**Recall:** If $Y \sim \mathrm{Bernoulli}(p)$, then

$$P(Y = y) = p^y(1-p)^{1-y}, \qquad y \in \{0, 1\}.$$

# Logistic regression: estimating the parameters

In logistic regression, we are assuming a model for $Y$. We typically estimate the parameter $\beta$ using maximum likelihood.

**Recall:** If $Y \sim \text{Bernoulli}(p)$, then

$$P(Y = y) = p^y (1-p)^{1-y}, \qquad y \in \{0, 1\}.$$

Thus, $L(p) = \prod_{i=1}^{n} p^{y_i} (1-p)^{1-y_i}$.

# Logistic regression: estimating the parameters

In logistic regression, we are assuming a model for $Y$. We typically estimate the parameter $\beta$ using maximum likelihood.

**Recall:** If $Y \sim \mathrm{Bernoulli}(p)$, then

$$P(Y = y) = p^y(1-p)^{1-y}, \qquad y \in \{0, 1\}.$$

Thus, $L(p) = \prod_{i=1}^n p^{y_i}(1-p)^{1-y_i}$.

Here $p = p(x_i, \beta) = \frac{e^{x_i^T \beta}}{1+e^{x_i^T \beta}}$. Therefore,

$$L(\beta) = \prod_{i=1}^n p(x_i, \beta)^{y_i}(1 - p(x_i, \beta))^{1-y_i}.$$

# Logistic regression: estimating the parameters

In logistic regression, we are assuming a model for $Y$. We typically estimate the parameter $\beta$ using maximum likelihood.

**Recall:** If $Y \sim \text{Bernoulli}(p)$, then

$$P(Y = y) = p^y(1-p)^{1-y}, \qquad y \in \{0,1\}.$$

Thus, $L(p) = \prod_{i=1}^n p^{y_i}(1-p)^{1-y_i}$.

Here $p = p(x_i, \beta) = \frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}$. Therefore,

$$L(\beta) = \prod_{i=1}^n p(x_i,\beta)^{y_i}(1-p(x_i,\beta))^{1-y_i}.$$

Taking the logarithm, we obtain

$$\begin{aligned} l(\beta) &= \sum_{i=1}^n y_i \log p(x_i,\beta) + (1-y_i)\log(1-p(x_i,\beta)) \\ &= \sum_{i=1}^n y_i(x_i^T\beta - \log(1+x_i^T\beta)) - (1-y_i)\log(1+e^{x_i^T\beta}) \\ &= \sum_{i=1}^n [y_i x_i^T\beta - \log(1+e^{x_i^T\beta})]. \end{aligned}$$

Taking the derivative:

$$\frac{\partial}{\partial \beta_j} l(\beta) = \sum_{i=1}^{n} \left[ y_i x_{ij} - x_{ij} \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}} \right].$$

Needs to be solved using numerical methods
(e.g. Newton-Raphson).

Logistic regression often performs well in applications.

As before, penalties can be added to regularize the problem or induce sparsity. For example,

$$\min_{\beta} -l(\beta) + \alpha \|\beta\|_1$$

$$\min_{\beta} -l(\beta) + \alpha \|\beta\|_2.$$

# Logistic regression with more than 2 classes

- Suppose now the response can take any of $\{1, \ldots, K\}$ values.
- Can still use logistic regression.
- We use the categorical distribution instead of the Bernoulli distribution.
- $P(Y = i|X = x) = p_i$, $0 \leq p_i \leq 1$, $\sum_{i=1}^{K} p_i = 1$.
- Each category has its own set of coefficients:

$$P(Y = i|X = x) = \frac{e^{x^T \beta^{(i)}}}{\sum_{i=1}^{K} e^{x^T \beta^{(i)}}}.$$

- Estimation can be done using maximum likelihood as for the binary case.

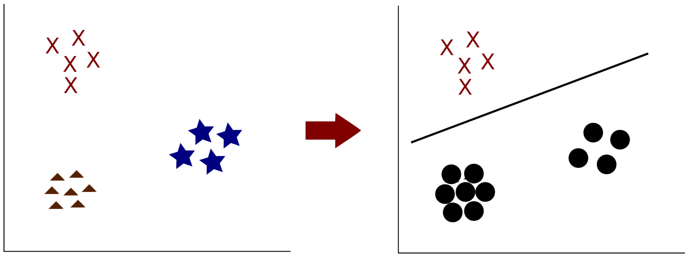
# Multiple classes of data

Other popular approaches to classify data from multiple categories.

# Multiple classes of data

Other popular approaches to classify data from multiple categories.

- **One versus all:**(or one versus the rest) Fit the model to separate each class against the remaining classes. Label a new point $x$ according to the model for which $x^T\beta + \beta_0$ is the largest.



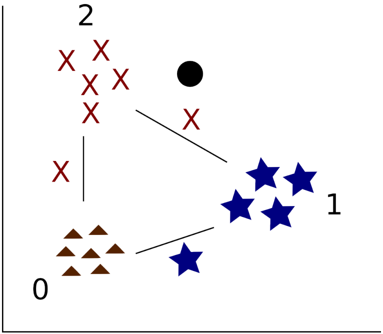
Need to fit the model $K$ times.

# Multiple classes of data (cont.)

- **One versus one:**
  1. Train a classifier for each possible **pair** of classes.
     Note: There are $\binom{K}{2} = K(K-1)/2$ such pairs.
  2. Classify a new points according to a **majority vote**: count the number of times the new point is assign to a given class, and pick the class with the largest number.

# Multiple classes of data (cont.)

- **One versus one:**
  1. Train a classifier for each possible **pair** of classes.
     Note: There are $\binom{K}{2} = K(K-1)/2$ such pairs.
  2. Classify a new points according to a **majority vote**: count the number of times the new point is assign to a given class, and pick the class with the largest number.




Need to fit the model $\binom{K}{2}$ times (computationally intensive).