Classification

December 20, 2017
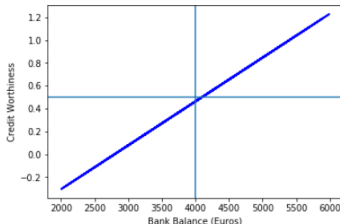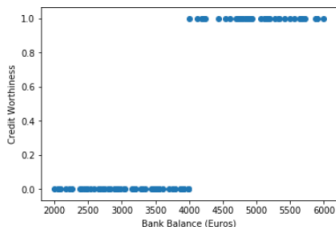
Classification

- Linear regression: the response variable y is quantitative.
- Classification: y is qualitative (takes on a number of discrete values).
- Classification problems seem to occur more often than regression problems:
    - spam classifiers (spam or ham)
    - classifying whether a bank transaction is fradulent or not
    - given a set of symptoms, determining which medical condition a person has
    - classifying whether a video is suitable or unsuitable for children
    - MNIST: given a handwritten digit, determine which digit it actually is

Why not Linear Regression?

Example 1

Consider the following (simplified) problem: given the bank balance x of an individual, determine whether they are credit worthy or not.
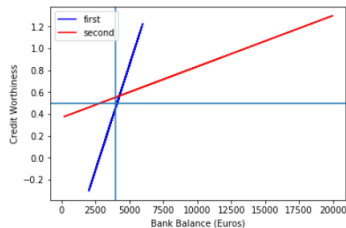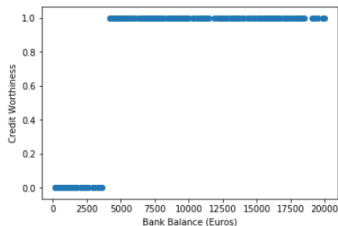

▶ Turns out that anyone with a balance of €4000 or more is credit worthy

▶ Classification: if $y(x) \geq 0.5$, then "credit worthy"; else "not"

▶ Slope of the regression line depends on the how many data points are in each of the two buckets

Example 1

▶ With more data points in the "positive" bucket, the slope of the regression line is less steep.

▶ The threshold predicted also changes.

Binary Classification: Logistic Regression

- Two classes: 0 and 1
- Logistic regression models the probability that the response variable y belongs to a particular class: $P(y = 1 \mid x)$
- $P(y = 1 \mid x; \theta) = g(\theta^\mathsf{T} x) = \frac{1}{1 + e^{-\theta^\mathsf{T} x}}$
- $g(z) = \frac{1}{1 + e^{-\theta^\mathsf{T} x}}$ is the sigmoid function

The Sigmoid Function

- $g(z) \to 1$ as $z \to \infty$
- $g(z) \to 0$ as $z \to -\infty$
- $g'(z) = g(z)(1 - g(z))$

In logistic regression, the probabilities are modeled as follows:

$$P(y = 1 \mid x; \theta) = g(\theta^\mathsf{T} x) = \frac{1}{1 + e^{-\theta^\mathsf{T} x}}$$

$$P(y = 0 \mid x; \theta) = 1 - g(\theta^\mathsf{T} x)$$

In logistic regression, the probabilities are modeled as follows:

$$P(y = 1 \mid x; \theta) = g(\theta^\mathsf{T} x) = \frac{1}{1 + e^{-\theta^\mathsf{T} x}}$$

$$P(y = 0 \mid x; \theta) = 1 - g(\theta^\mathsf{T} x)$$

$$P(y \mid x; \theta) = (g(\theta^\mathsf{T} x))^y \cdot (1 - g(\theta^\mathsf{T} x))^{1-y}$$

In logistic regression, the probabilities are modeled as follows:

$$P(y = 1 \mid x; \theta) = g(\theta^\mathsf{T} x) = \frac{1}{1 + e^{-\theta^\mathsf{T} x}}$$

$$P(y = 0 \mid x; \theta) = 1 - g(\theta^\mathsf{T} x)$$

$$P(y \mid x; \theta) = (g(\theta^\mathsf{T} x))^y \cdot (1 - g(\theta^\mathsf{T} x))^{1-y}$$

The Likelihood Function

Assume that the m training examples $(x^{(1)}, y^{(1)}), \ldots (x^{(m)}, y^{(m)})$ were generated *independently*.

$$L(\theta; \{(x^{(i)}, y^{(i)})\}_{i=1}^{m}) = \prod_{i=1}^{m} P(y^{(i)} \mid x^{(i)}; \theta)$$

$$= \prod_{i=1}^{m} \left(g(\theta^\mathsf{T} x^{(i)}) \right)^{y^{(i)}} \cdot \left(1 - g(\theta^\mathsf{T} x^{(i)}) \right)^{(1-y^{(i)})}$$

. . . is equivalent to maximizing the log-likelihood:

$$l(\theta) = \log L(\theta)$$
$$= \sum_{i=1}^{m} y^{(i)} \log g(\theta^{\mathsf{T}} x^{(i)}) + (1 - y^{(i)}) \log(1 - g(\theta^{\mathsf{T}} \cdot x^{(i)}))$$

- ▶ Use gradient ascent to maximize $l(\theta)$
- ▶ $\theta_{\mathsf{new}} := \theta_{\mathsf{old}} + \alpha \nabla_\theta l(\theta)$
- ▶ $\theta_j := \theta_j + \alpha \cdot \sum_{i=1}^{m} \left(y^{(i)} - g(\theta^{\mathsf{T}} x^{(i)}) \right) \cdot x_j^{(i)}$
- ▶ Stochastic gradient ascent: $\theta_j := \theta_j + \alpha \cdot (y^{(i)} - g(\theta^{\mathsf{T}} \cdot x^{(i)})) \cdot x_j^{(i)}$

Classifying New Data Points

▶ Once θ has been estimated (using maximum likelihood, for instance), given x, $P(y = 1 \mid x) = g(\theta^\mathsf{T} \cdot x)$

▶ We could for instance classify x as belonging to class 1 iff $g(\theta^\mathsf{T} \cdot x) \geq 0.5$

Harder than evaluating a linear regressor.

▶ Consider unbalanced data sets: let's say that 90% of customers in an online store are one-time customers.

▶ Want to determine whether a customer is a one-time customer.

▶ A "dumb" classifier that declares every customer as "bad" has 90% accuracy.

Precision and Recall

- Accuracy of positive predictions = Precision = $\frac{\text{TP}}{\text{TP+FP}}$
- Fraction of positive instances correctly classified = Recall = $\frac{\text{TP}}{\text{TP+FN}}$

Stochastic Gradient Descent

In batch gradient descent, the update step for the jth component is:

$$\theta_j := \theta_j + \alpha \cdot \sum_{i=1}^{m} \left(y^{(i)} - \sum_{j=0}^{n} x_j^{(i)} \theta_j \right) \cdot x_j^{(i)}.$$

▶ Has to scan through the entire training set for a *single* update
▶ Costly operation if m is large

Stochastic Gradient Descent

In batch gradient descent, the update step for the jth component is:

$$\theta_j := \theta_j + \alpha \cdot \sum_{i=1}^{m} \left(y^{(i)} - \sum_{j=0}^{n} x_j^{(i)} \theta_j \right) \cdot x_j^{(i)}.$$

- ▶ Has to scan through the entire training set for a *single* update
- ▶ Costly operation if m is large
- ▶ **Stochastic Gradient Descent**: for every training instance (x, y), $x = (x_0, x_1, \ldots, x_n)^{\mathsf{T}}$, update the parameters:

$$\theta_j := \theta_j + \alpha \cdot \left(y - \sum_{j=0}^{n} x_j \theta_j \right) \cdot x_j.$$

Stochastic Gradient Descent: Features and Issues

▶ Doesn't have to look at the entire training set to make progress.

▶ Often gets close to the optimum much faster than batch gradient descent.

▶ May never converge to the optimum (can keep on oscillating between values near the optimum). This problem is alleviated by choosing α to be very small.

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{m} \left(y^{(i)} - \sum_{j=0}^{n} x_j^{(i)} \theta_j \right)^2$$

Want to find in closed-form a value of θ that minimizes $J(\theta)$

▶ Write $J(\theta)$ is matrix-vector form.

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{m} \left(y^{(i)} - \sum_{j=0}^{n} x_j^{(i)} \theta_j \right)^2$$

Want to find in closed-form a value of θ that minimizes $J(\theta)$

▶ Write $J(\theta)$ is matrix-vector form.

▶ **Design matrix.** An $m \times (n+1)$ matrix X defined by:

$$X = \begin{bmatrix} — & (x^{(1)})^{\mathsf{T}} & — \\ — & (x^{(2)})^{\mathsf{T}} & — \\ \vdots & \vdots & \vdots \\ — & (x^{(m)})^{\mathsf{T}} & — \end{bmatrix}$$

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{m} \left(y^{(i)} - \sum_{j=0}^{n} x_j^{(i)} \theta_j \right)^2$$

Want to find in closed-form a value of θ that minimizes $J(\theta)$

- ▶ Write $J(\theta)$ is matrix-vector form.
- ▶ **Design matrix.** An $m \times (n+1)$ matrix X defined by:

$$X = \begin{bmatrix} — & (x^{(1)})^{\mathsf{T}} & — \\ — & (x^{(2)})^{\mathsf{T}} & — \\ \vdots & \vdots & \vdots \\ — & (x^{(m)})^{\mathsf{T}} & — \end{bmatrix}$$

- ▶ $y = (y^{(1)}, \ldots, y^{(m)})^{\mathsf{T}}$

$$y - X \cdot \theta = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix} - \begin{bmatrix} (x^{(1)})^\intercal \cdot \theta \\ (x^{(2)})^\intercal \cdot \theta \\ \vdots \\ (x^{(m)})^\intercal \cdot \theta \end{bmatrix}$$

Matrix-form of $J(\theta)$:

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{m} \left(y^{(i)} - \sum_{j=0}^{n} x_j^{(i)} \theta_j \right)^2$$

$$= \frac{1}{2} (y - X\theta)^\intercal (y - X\theta)$$

Minimize $J(\theta)$ w.r.t θ:

$$\nabla_\theta J(\theta) = \nabla_\theta \frac{1}{2}(y - X\theta)^\mathsf{T}(y - X\theta)$$
$$= \text{see Andrew Ng's notes}$$
$$= X^\mathsf{T} X\theta - X^\mathsf{T} y$$

yielding:

$$\boxed{\theta = (X^\mathsf{T} X)^{-1} \cdot X^\mathsf{T} y.}$$

Minimize $J(\theta)$ w.r.t θ:

$$\nabla_\theta J(\theta) = \nabla_\theta \frac{1}{2}(y - X\theta)^\intercal (y - X\theta)$$
$$= \text{see Andrew Ng's notes}$$
$$= X^\intercal X\theta - X^\intercal y$$

yielding:

$$\boxed{\theta = (X^\intercal X)^{-1} \cdot X^\intercal y.}$$

▶ **Assumption:** X has full column rank so that $X^\intercal X$ is invertible.

Minimize $J(\theta)$ w.r.t θ:

$$\nabla_\theta J(\theta) = \nabla_\theta \frac{1}{2}(y - X\theta)^\intercal(y - X\theta)$$
$$= \text{see Andrew Ng's notes}$$
$$= X^\intercal X\theta - X^\intercal y$$

yielding:

$$\boxed{\theta = (X^\intercal X)^{-1} \cdot X^\intercal y.}$$

▶ **Assumption:** X has full column rank so that $X^\intercal X$ is invertible.
Proof. Show that $X^\intercal X z = 0$ implies $z = 0$.

Minimize $J(\theta)$ w.r.t θ:

$$\nabla_\theta J(\theta) = \nabla_\theta \frac{1}{2}(y - X\theta)^\intercal(y - X\theta)$$
$$= \text{see Andrew Ng's notes}$$
$$= X^\intercal X\theta - X^\intercal y$$

yielding:

$$\boxed{\theta = (X^\intercal X)^{-1} \cdot X^\intercal y.}$$

- **Assumption:** X has full column rank so that $X^\intercal X$ is invertible.
  Proof. Show that $X^\intercal X z = 0$ implies $z = 0$.
- If X does not have full column rank, the usual strategy is to remove redundant columns.

Assumptions

$$y^{(i)} = \theta^\mathsf{T} \cdot x^{(i)} + \epsilon^{(i)},$$

where the error terms $\epsilon^{(i)}$

▶ capture unmodeled effects and/or random noise

▶ are independent and identically distributed as $N(0, \sigma^2)$

Assumptions

$$y^{(i)} = \theta^{\mathsf{T}} \cdot x^{(i)} + \epsilon^{(i)},$$

where the error terms $\epsilon^{(i)}$

▶ capture unmodeled effects and/or random noise

▶ are independent and identically distributed as $N(0, \sigma^2)$

Given $x^{(i)}$,

▶ $E(y^{(i)} \mid x^{(i)}) = \theta^{\mathsf{T}} \cdot x^{(i)}$

Probabilistic Interpretation

Assumptions

$$y^{(i)} = \theta^\mathsf{T} \cdot x^{(i)} + \epsilon^{(i)},$$

where the error terms $\epsilon^{(i)}$

- ▶ capture unmodeled effects and/or random noise
- ▶ are independent and identically distributed as $N(0, \sigma^2)$

Given $x^{(i)}$,

- ▶ $E(y^{(i)} \mid x^{(i)}) = \theta^\mathsf{T} \cdot x^{(i)}$
- ▶ $\mathrm{Var}(y^{(i)} \mid x^{(i)}) = \mathrm{Var}(\epsilon^{(i)}) = \sigma^2$

Probabilistic Interpretation

Assumptions

$$y^{(i)} = \theta^\mathsf{T} \cdot x^{(i)} + \epsilon^{(i)},$$

where the error terms $\epsilon^{(i)}$

- ▶ capture unmodeled effects and/or random noise
- ▶ are independent and identically distributed as $N(0, \sigma^2)$

Given $x^{(i)}$,

- ▶ $E(y^{(i)} \mid x^{(i)}) = \theta^\mathsf{T} \cdot x^{(i)}$
- ▶ $\mathrm{Var}(y^{(i)} \mid x^{(i)}) = \mathrm{Var}(\epsilon^{(i)}) = \sigma^2$

Thus $y^{(i)} \mid x^{(i)} \sim N(\theta^\mathsf{T} x^{(i)}, \sigma^2)$:

$$p(y^{(i)} \mid x^{(i)}; \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^\mathsf{T} \cdot x^{(i)})^2}{2\sigma^2}\right)$$

▶ Given $x = (x^{(1)}, \ldots, x^{(m)})^{\mathsf{T}}$ and θ, what is the joint distribution of the $y = (y^{(1)}, \ldots, y^{(m)})^{\mathsf{T}}$?

Maximum Likelihood Estimation

▶ Given $x = (x^{(1)}, \ldots, x^{(m)})^{\mathsf{T}}$ and θ, what is the joint distribution of the $y = (y^{(1)}, \ldots, y^{(m)})^{\mathsf{T}}$?

▶ Since the $\epsilon^{(i)}$s are independent:

▶ Given $x = (x^{(1)}, \ldots, x^{(m)})^\mathsf{T}$ and θ, what is the joint distribution of the $y = (y^{(1)}, \ldots, y^{(m)})^\mathsf{T}$?

▶ Since the $\epsilon^{(i)}$s are independent:

$$p(y \mid x; \theta) = \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}; \theta)$$

Maximum Likelihood Estimation

▶ Given $x = (x^{(1)}, \ldots, x^{(m)})^\mathsf{T}$ and θ, what is the joint distribution of the $y = (y^{(1)}, \ldots, y^{(m)})^\mathsf{T}$?

▶ Since the $\epsilon^{(i)}$s are independent:

$$p(y \mid x; \theta) = \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}; \theta)$$

▶ **Likelihood function.** $L(\theta) = p(y \mid x; \theta)$

Maximum Likelihood Estimation

▶ Given $x = (x^{(1)}, \ldots, x^{(m)})^{\mathsf{T}}$ and θ, what is the joint distribution of the $y = (y^{(1)}, \ldots, y^{(m)})^{\mathsf{T}}$?

▶ Since the $\epsilon^{(i)}$s are independent:

$$p(y \mid x; \theta) = \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}; \theta)$$

▶ **Likelihood function.** $L(\theta) = p(y \mid x; \theta)$

$$L(\theta) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{\left(y^{(i)} - \theta^{\mathsf{T}} x^{(i)}\right)^2}{2\sigma^2} \right\}$$

Maximum Likelihood Estimation

▶ Given $x = (x^{(1)}, \ldots, x^{(m)})^{\mathsf{T}}$ and θ, what is the joint distribution of the $y = (y^{(1)}, \ldots, y^{(m)})^{\mathsf{T}}$?

▶ Since the $\epsilon^{(i)}$s are independent:

$$p(y \mid x; \theta) = \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}; \theta)$$

▶ **Likelihood function.** $L(\theta) = p(y \mid x; \theta)$

$$L(\theta) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{\left(y^{(i)} - \theta^{\mathsf{T}} x^{(i)}\right)^2}{2\sigma^2} \right\}$$

▶ **Principle of Maximum Likelihood.** Choose the parameters to make the data as likely as possible. Choose θ to maximize $L(\theta)$.

Maximum Likelihood Estimation . . .

Maximizing $L(\theta)$ is equivalent to maximizing *any* strictly increasing function of $L(\theta)$.

Maximizing $L(\theta)$ is equivalent to maximizing *any* strictly increasing function of $L(\theta)$.

- ▶ Usual to maximize the log likelihood $l(\theta) = \log L(\theta)$.
- ▶ $l(\theta) = m \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2} \cdot \frac{1}{2} \sum_{i=1}^{m} \left(y^{(i)} - \theta^{\intercal} x^{(i)} \right)^2$.
- ▶ Maximizing $l(\theta)$ is equivalent to minimizing $\frac{1}{2} \sum_{i=1}^{m} \left(y^{(i)} - \theta^{\intercal} x^{(i)} \right)^2$.

Maximizing $L(\theta)$ is equivalent to maximizing *any* strictly increasing function of $L(\theta)$.

- Usual to maximize the log likelihood $l(\theta) = \log L(\theta)$.
- $l(\theta) = m \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2} \cdot \frac{1}{2} \sum_{i=1}^{m} \left(y^{(i)} - \theta^{\mathsf{T}} x^{(i)} \right)^2$.
- Maximizing $l(\theta)$ is equivalent to minimizing $\frac{1}{2} \sum_{i=1}^{m} \left(y^{(i)} - \theta^{\mathsf{T}} x^{(i)} \right)^2$.

Summary

Under the previous probabilistic assumptions: **least-squares regression** corresponds to finding the **maximum likelihood estimate** of θ.

▶ Residual Standard Error: standard deviation of the error terms $\epsilon^{(i)}$.

$$\text{RSE} = \sqrt{\frac{1}{n-2} \sum_{i=1}^{m} \left(y^{(i)} - \hat{y}^{(i)}\right)^2}$$

The Goodness of Fit

▶ Residual Standard Error: standard deviation of the error terms $\epsilon^{(i)}$.

$$\text{RSE} = \sqrt{\frac{1}{n-2} \sum_{i=1}^{m} \left(y^{(i)} - \hat{y}^{(i)}\right)^2}$$

▶ R^2 Statistic:

$$R^2 = 1 - \frac{\sum_{i=1}^{m} \left(y^{(i)} - \hat{y}^{(i)}\right)^2}{\sum_{i=1}^{m} \left(y^{(i)} - \bar{y}\right)^2}$$

The Goodness of Fit

▶ Residual Standard Error: standard deviation of the error terms $\epsilon^{(i)}$.

$$\text{RSE} = \sqrt{\frac{1}{n-2} \sum_{i=1}^{m} \left(y^{(i)} - \hat{y}^{(i)}\right)^2}$$

▶ R^2 Statistic:

$$R^2 = 1 - \frac{\sum_{i=1}^{m} \left(y^{(i)} - \hat{y}^{(i)}\right)^2}{\sum_{i=1}^{m} \left(y^{(i)} - \bar{y}\right)^2}$$

▶ Total sum of squares $= \sum_{i=1}^{m} \left(y^{(i)} - \bar{y}\right)^2$: variability inherent in the response

▶ Residual sum of squares $= \left(y^{(i)} - \hat{y}^{(i)}\right)^2$: variability left unexplained after performing the regression