Classification

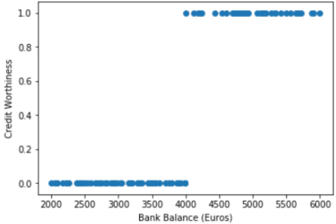
December 17, 2017

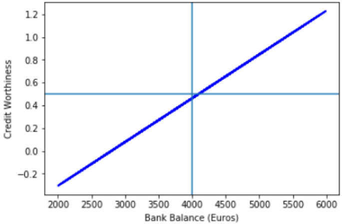
Classification

- 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

Example 1

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



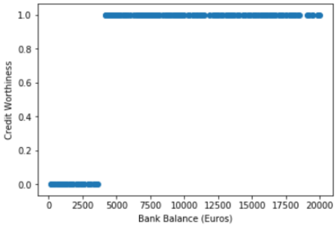


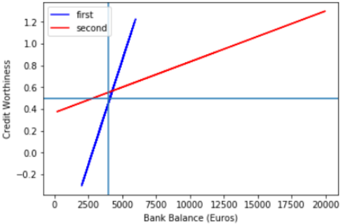
- Turns out that anyone with a balance of €4000 or more is credit worthy
- Classification: if $y(x) \geq 0.5$, then "credit worthy"; else "not"
- Slope of the regression line depends on the how many data points are in each of the two buckets

Why not Linear Regression?

Example 1

- With more data points in the "positive" bucket, the slope of the regression line is less steep.
- The threshold predicted also changes.





Linear Regression Models and Least Squares ...

- Typical situation: we have training data

$$(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})$$

  from which to estimate the parameters θ.

- Least squares: pick parameters $\theta = (\theta_0, \ldots, \theta_n)^\mathsf{T}$ to minimize

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

Linear Regression Models and Least Squares . . .

- Typical situation: we have training data

$$(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})$$

  from which to estimate the parameters θ.

- Least squares: pick parameters $\theta = (\theta_0, \ldots, \theta_n)^{\mathsf{T}}$ to minimize

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

- Gradient descent
- Analytical solution
- Probabilistic interpretation

Gradient Descent

- Start with an “initial guess” for θ

Gradient Descent

- Start with an "initial guess" for θ
- Repeatedly perform the update for all $0 \leq j \leq n$:

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta),$$

where α is the learning rate.

Gradient Descent

- Start with an "initial guess" for θ
- Repeatedly perform the update for all $0 \leq j \leq n$:

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta),$$

  where α is the learning rate.
- For each $0 \leq j \leq n$: $\partial J(\theta)/\partial \theta_j = -\sum_{i=1}^{m} \left(y^{(i)} - \sum_{j=0}^{n} x_j^{(i)} \theta_j \right) \cdot x_j^{(i)}$.

Gradient Descent

- Start with an "initial guess" for θ
- Repeatedly perform the update for all $0 \leq j \leq n$:

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta),$$

  where α is the learning rate.
- For each $0 \leq j \leq n$: $\partial J(\theta)/\partial \theta_j = -\sum_{i=1}^{m} \left(y^{(i)} - \sum_{j=0}^{n} x_j^{(i)} \theta_j \right) \cdot x_j^{(i)}$.
- Magnitude of update: a linear function of the input vectors, where the coefficients are the error terms $y - \sum_j x_j \theta_j$

Gradient Descent

- Start with an "initial guess" for θ
- Repeatedly perform the update for all $0 \le j \le n$:

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta),$$

  where α is the learning rate.
- For each $0 \le j \le n$: $\partial J(\theta)/\partial \theta_j = -\sum_{i=1}^{m} \left(y^{(i)} - \sum_{j=0}^{n} x_j^{(i)} \theta_j \right) \cdot x_j^{(i)}$.
- Magnitude of update: a linear function of the input vectors, where the coefficients are the error terms $y - \sum_j x_j \theta_j$
- Looks at every input in the training set before making an update:

Gradient Descent

- Start with an "initial guess" for θ
- Repeatedly perform the update for all $0 \leq j \leq n$:

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta),$$

  where α is the learning rate.
- For each $0 \leq j \leq n$: $\partial J(\theta)/\partial\theta_j = -\sum_{i=1}^{m} \left(y^{(i)} - \sum_{j=0}^{n} x_j^{(i)} \theta_j \right) \cdot x_j^{(i)}$.
- Magnitude of update: a linear function of the input vectors, where the coefficients are the error terms $y - \sum_j x_j \theta_j$
- Looks at every input in the training set before making an update: **batch gradient descent**

Gradient Descent

- Start with an "initial guess" for θ
- Repeatedly perform the update for all $0 \leq j \leq n$:

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta),$$

  where α is the learning rate.
- For each $0 \leq j \leq n$: $\partial J(\theta)/\partial \theta_j = -\sum_{i=1}^{m} \left(y^{(i)} - \sum_{j=0}^{n} x_j^{(i)} \theta_j \right) \cdot x_j^{(i)}$.
- Magnitude of update: a linear function of the input vectors, where the coefficients are the error terms $y - \sum_j x_j \theta_j$
- Looks at every input in the training set before making an update: **batch gradient descent**
- Gradient descent is susceptible to **local minima** in general; in this case, J is a **convex** function and has a **unique global minimum**.

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.

Analytic Solution

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

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

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

Analytic Solution

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

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:

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

Analytic Solution

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

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:

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

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

Analytic Solution ...

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

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

$$\begin{aligned} J(\theta) &= \frac{1}{2} \sum_{i=1}^{m} \left(y^{(i)} - \sum_{j=0}^{n} x_j^{(i)} \theta_j \right)^2 \\ &= \frac{1}{2} (y - X\theta)^\mathsf{T} (y - X\theta) \end{aligned}$$

Analytic Solution . . .

Minimize $J(\theta)$ w.r.t θ:

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

yielding:

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

Analytic Solution . . .

Minimize $J(\theta)$ w.r.t θ:

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

yielding:

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

- **Assumption:** X has full column rank so that $X^{\mathsf{T}}X$ is invertible.

Analytic Solution ...

Minimize $J(\theta)$ w.r.t θ:

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

yielding:

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

- **Assumption:** X has full column rank so that $X^\mathsf{T} X$ is invertible.
  Proof. Show that $X^\mathsf{T} X z = 0$ implies $z = 0$.

Analytic Solution . . .

Minimize $J(\theta)$ w.r.t θ:

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

yielding:

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

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

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 $N(0, \sigma^2)$

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 $x^{(i)}$,

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

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

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

Maximum Likelihood Estimation

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

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:

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:

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

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 $L(\theta)$.

Maximum Likelihood Estimation ...

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 $\frac{1}{2} \sum_{i=1}^{m} \left(y^{(i)} - \theta^{\mathsf{T}} x^{(i)}\right)^2$.

Maximum Likelihood Estimation . . .

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.

The Goodness of Fit

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

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

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 $\epsilon^{(i)}$.

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

- 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