Anomaly Detection
Linear models - leverage scores

Paul Irofti
Cristian Rusu
Andrei Pătraşcu

Computer Science Department
University of Bucharest

Class structure

Topics for today:

- review of the Gaussian pdf
- review of least-squares
- leverage scores definition and properties
- leverage scores for anomaly detection

Gaussian pdf

one dimensional Gaussian

$$\mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

multi-dimensional Gaussian

$$\mathcal{N}(\mu, \Sigma) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\{-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)\}$$

In general, we say that we sample from a standard Gaussian variable:

$x \sim \mathcal{N}(0, 1)$ or $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$

Note: there is already a hint that “Σ is the square of something” statistical variables have two important properties:

the mean of the variable: $\mathbb{E}[\mathbf{x}] = \mu$

the variance of the variable: $\mathbb{E}[(\mathbf{x} - \mathbb{E}[x])(\mathbf{x} - \mathbb{E}[x])^T] = \Sigma$

An exercise for you: you are in the one-dimensional setting and you have a Gaussian variable $x \sim \mathcal{N}(\mu, \sigma^2)$ and then we need to build a new variable $y = ax + b$. what sort of random variable is this?

An exercise for you: you are in the one-dimensional setting and you have a Gaussian variable $x \sim \mathcal{N}(\mu, \sigma^2)$ and then we need to build a new variable $y = ax + b$. what sort of random variable is this?

$\mathbb{E}[y] = \mathbb{E}[ax + b] = a\mu + b$

$\mathbb{E}[(y - \mathbb{E})(y - \mathbb{E})^T] = a^2\mathbb{E}[(x - \mu)(x - \mu)] = a^2\sigma^2$
where we use the fact that $\mathbb{E}[y - \mathbb{E}[y]] = a\mathbb{E}[X - \mu]$

Another exercise for you: you are given a one-dimensional standard Gaussian variable $x \sim \mathcal{N}(0, 1)$, how do you convert it into another standard Gaussian variable with mean μ and variance σ^2?

Another exercise for you: you are given a one-dimensional standard Gaussian variable $x \sim \mathcal{N}(0, 1)$, how do you convert it into another standard Gaussian variable with mean μ and variance σ^2?

$$y = \mu + \sigma x$$

What would be the reverse of this?

Another exercise for you: you are given a one-dimensional standard Gaussian variable $x \sim \mathcal{N}(0, 1)$, how do you convert it into another standard Gaussian variable with mean μ and variance σ^2?

$y = \mu + \sigma x$

What would be the reverse of this?

$y = \frac{x - \mu}{\sigma}$ (we standardize the random variable)

Another exercise for you: you are given a d-dimensional standard Gaussian variable $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, how do you convert it into another standard Gaussian variable with mean μ and variance Σ?

Another exercise for you: you are given a d-dimensional standard Gaussian variable $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, how do you convert it into another standard Gaussian variable with mean μ and variance Σ?

$\mathbf{y} = \mu + \mathbf{L}\mathbf{x}$ where $LL^T = \Sigma$ (from the Cholesky factorization of Σ, this is the "square root" for a matrix).

the setup in this class is the following:

- we are in the supervised setting
- we are given a dataset where each data point has d features
- we are given n data points $\mathbf{x}_i \in \mathbb{R}^d$, the features
- we are given n labels for these data points $y_i \in \mathbb{R}$

the goals are:

- assume a linear predictor $\beta \in \mathbb{R}^d$
- estimate the best linear predictor from the data, i.e., $\mathbf{x}_i^T \beta \approx y_i$ for all $i = 1, \ldots, n$
- pick the squared error to minimize $(\mathbf{x}_i^T \beta - y_i)^2$ for all $i = 1, \ldots, n$
- overall objective function is $\sum_{i=1}^n (\mathbf{x}_i^T \beta - y_i)^2$

Least-squares

- overall objective function is:

$$\sum_{i=1}^{n}(\mathbf{x}_i^T\beta - y_i)^2 \tag{1}$$

- this can be written in matrix form as:

$$\|\mathbf{X}\beta - \mathbf{y}\|_F^2 \tag{2}$$

- $\mathbf{X}$ is an $n \times d$ matrix where the i^{th} row is $\mathbf{x}_i^T$
- $\mathbf{y}$ is an n-dimensional vector of labels
- the unknown is β the d-dimensional vector
- we have used the Frobenius norm $\|\mathbf{A}\|_F^2 = \mathrm{tr}(\mathbf{A}^T\mathbf{A}) = \sum_{i=1}^{n}\sum_{j=1}^{d} = |A_{ij}|^2$, for vectors this is just $\|\mathbf{x}\|_F^2 = \mathbf{x}^T\mathbf{x} = \sum_{i=1}^{n}|x_i|^2 = \|\mathbf{x}\|_2^2$.

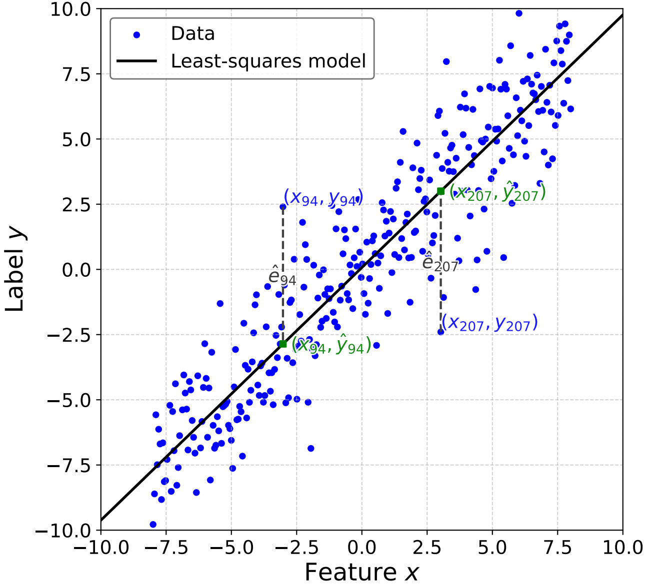
The least-squares problem solves the following:

$$\underset{\beta}{\text{minimize}} \ \|\mathbf{X}\beta - \mathbf{y}\|_F^2 \tag{3}$$

- when $n = d$ we have $\beta^\star = \mathbf{X}^{-1}\mathbf{y}$
- when $n > d$ we have $\beta^\star = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$
- when $n < d$ we have $\beta^\star = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{y}$

- how do we get these?
- what happens if we replace the squared with absolute value?
- how do we compute $\beta^\star$ in each case above?

Least-squares problems



There are several things that the least-squares assumes:

- we assume that the data was generated as $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$ where $\mathbf{e}$ is considered to be a standard Gaussian random variable: $\mathbb{E}[\mathbf{e}] = \mathbf{0}$ and $\text{var}[\mathbf{e}] = \sigma^2 \mathbf{I}_n$
- note that $\mathbb{E}[\mathbf{y}] = \mathbf{X}\beta$ and $\text{var}[\mathbf{y}] = \sigma^2 \mathbf{I}_n$
- the the least-squares solution is given by $\beta^\star = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$
- the projected values are given by $\hat{\mathbf{y}} = \mathbf{X}\beta^\star = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{H}\mathbf{y}$
- and the the empirical error is given by $\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$ where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$

A few properties for **H**:

- **H** is symmetric.

  Proof.

A few properties for $\mathbf{H}$:

- $\mathbf{H}$ is symmetric.
  Proof. Check that $\mathbf{H} = \mathbf{H}^T$.
- $\mathbf{H}$ is positive semi-definite.
  Proof.

A few properties for $\mathbf{H}$:

- $\mathbf{H}$ is symmetric.

  Proof. Check that $\mathbf{H} = \mathbf{H}^T$.

- $\mathbf{H}$ is positive semi-definite.

  Proof. True because $(\mathbf{X}^T\mathbf{X})^{-1}$ is positive definite.

- $\mathbf{H}^2 = \mathbf{H}$.

  Proof.

A few properties for $\mathbf{H}$:

- $\mathbf{H}$ is symmetric.
  Proof. Check that $\mathbf{H} = \mathbf{H}^T$.
- $\mathbf{H}$ is positive semi-definite.
  Proof. True because $(\mathbf{X}^T\mathbf{X})^{-1}$ is positive definite.
- $\mathbf{H}^2 = \mathbf{H}$.
  Proof. Use the definition and simplify the expression.
- $(\mathbf{I}_n - \mathbf{H})^2 = (\mathbf{I}_n - \mathbf{H})$.
  Proof.

A few properties for $\mathbf{H}$:

- $\mathbf{H}$ is symmetric.

  Proof. Check that $\mathbf{H} = \mathbf{H}^T$.

- $\mathbf{H}$ is positive semi-definite.

  Proof. True because $(\mathbf{X}^T\mathbf{X})^{-1}$ is positive definite.

- $\mathbf{H}^2 = \mathbf{H}$.

  Proof. Use the definition and simplify the expression.

- $(\mathbf{I}_n - \mathbf{H})^2 = (\mathbf{I}_n - \mathbf{H})$.

  Proof. Use the definition and square the quantity explicitly.

- $\mathrm{tr}(\mathbf{H}) = d$

  Proof.

A few properties for $\mathbf{H}$:

- $\mathbf{H}$ is symmetric.
  Proof. Check that $\mathbf{H} = \mathbf{H}^T$.
- $\mathbf{H}$ is positive semi-definite.
  Proof. True because $(\mathbf{X}^T\mathbf{X})^{-1}$ is positive definite.
- $\mathbf{H}^2 = \mathbf{H}$.
  Proof. Use the definition and simplify the expression.
- $(\mathbf{I}_n - \mathbf{H})^2 = (\mathbf{I}_n - \mathbf{H})$.
  Proof. Use the definition and square the quantity explicitly.
- $\mathrm{tr}(\mathbf{H}) = d$
  Proof. $\mathrm{tr}(\mathbf{H}) = \mathrm{tr}(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T) = \mathrm{tr}(\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}) = \mathrm{tr}(\mathbf{I}_d) = d$.

There are several things that the least-squares assumes:

- we assume that the data was generated as $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$ where $\mathbf{e}$ is considered to be a standard Gaussian random variable: $\mathbb{E}[\mathbf{e}] = \mathbf{0}$ and $\text{var}[\mathbf{e}] = \sigma^2\mathbf{I}_n$
- note that $\mathbb{E}[\mathbf{y}] = \mathbf{X}\beta$ and $\text{var}[\mathbf{y}] = \sigma^2\mathbf{I}_n$
- the the least-squares solution is given by $\beta^\star = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$
- the projected values are given by $\hat{\mathbf{y}} = \mathbf{X}\beta^\star = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{H}\mathbf{y}$
- and the the empirical error is given by $\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$ where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$
- then, $\mathbb{E}[\hat{\mathbf{y}}] =$

There are several things that the least-squares assumes:

- we assume that the data was generated as $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$ where $\mathbf{e}$ is considered to be a standard Gaussian random variable: $\mathbb{E}[\mathbf{e}] = \mathbf{0}$ and $\mathrm{var}[\mathbf{e}] = \sigma^2\mathbf{I}_n$
- note that $\mathbb{E}[\mathbf{y}] = \mathbf{X}\beta$ and $\mathrm{var}[\mathbf{y}] = \sigma^2\mathbf{I}_n$
- the the least-squares solution is given by $\beta^\star = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$
- the projected values are given by $\hat{\mathbf{y}} = \mathbf{X}\beta^\star = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{H}\mathbf{y}$
- and the the empirical error is given by $\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$ where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$
- then, $\mathbb{E}[\hat{\mathbf{y}}] = \mathbf{X}\beta$, $\mathrm{var}[\hat{\mathbf{y}}] =$

There are several things that the least-squares assumes:

- we assume that the data was generated as $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$ where $\mathbf{e}$ is considered to be a standard Gaussian random variable: $\mathbb{E}[\mathbf{e}] = \mathbf{0}$ and $\text{var}[\mathbf{e}] = \sigma^2 \mathbf{I}_n$
- note that $\mathbb{E}[\mathbf{y}] = \mathbf{X}\beta$ and $\text{var}[\mathbf{y}] = \sigma^2 \mathbf{I}_n$
- the the least-squares solution is given by $\beta^\star = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$
- the projected values are given by $\hat{\mathbf{y}} = \mathbf{X}\beta^\star = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{H}\mathbf{y}$
- and the the empirical error is given by $\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$ where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$
- then, $\mathbb{E}[\hat{\mathbf{y}}] = \mathbf{X}\beta$, $\text{var}[\hat{\mathbf{y}}] = \sigma^2\mathbf{H}^2 = \sigma^2\mathbf{H}$ and $\text{var}[\hat{\mathbf{e}}] =$

There are several things that the least-squares assumes:

- we assume that the data was generated as $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$ where $\mathbf{e}$ is considered to be a standard Gaussian random variable: $\mathbb{E}[\mathbf{e}] = \mathbf{0}$ and $\text{var}[\mathbf{e}] = \sigma^2 \mathbf{I}_n$
- note that $\mathbb{E}[\mathbf{y}] = \mathbf{X}\beta$ and $\text{var}[\mathbf{y}] = \sigma^2 \mathbf{I}_n$
- the the least-squares solution is given by $\beta^\star = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$
- the projected values are given by $\hat{\mathbf{y}} = \mathbf{X}\beta^\star = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{H}\mathbf{y}$
- and the the empirical error is given by $\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$ where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$
- then, $\mathbb{E}[\hat{\mathbf{y}}] = \mathbf{X}\beta$, $\text{var}[\hat{\mathbf{y}}] = \sigma^2\mathbf{H}^2 = \sigma^2\mathbf{H}$ and $\text{var}[\hat{\mathbf{e}}] = \sigma^2(\mathbf{I}_n - \mathbf{H})^2 = \sigma^2(\mathbf{I}_n - \mathbf{H})$

The leverage scores are the diagonal elements of the $\mathbf{H}$ matrix, i.e., $h_i = H_{ii} = \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i$.

We have the following properties:

- $0 \leq h_i \leq 1$.
- $\sum_{i=1}^n h_i = d$.
  Proof.

The leverage scores are the diagonal elements of the **H** matrix, i.e., $h_i = H_{ii} = \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i$.

We have the following properties:

- $0 \leq h_i \leq 1$.
- $\sum_{i=1}^n h_i = d$.
  Proof. The diagonal of **H** has only positive entries that sum up to d.

Why are these scores so important? They show the self-sensitivity of each residual:

$$h_{ii} = \frac{\partial \hat{y}_i}{\partial y_i} \tag{4}$$

This measures the degree by which the i^{th} measured value y_i influences the i^{th} predicted value $\hat{y}_i$.

What are considered high values? Those who deviate a lot from the expected value of the leverage scores. What is this value?

Why are these scores so important? They show the self-sensitivity of each residual:

$$h_{ii} = \frac{\partial \hat{y}_i}{\partial y_i} \tag{4}$$

This measures the degree by which the i^{th} measured value y_i influences the i^{th} predicted value $\hat{y}_i$.

What are considered high values? Those who deviate a lot from the expected value of the leverage scores. What is this value? $\bar{h} = \frac{d}{n}$.

A score, similar to the z-score we have talked about in the past:

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}} \tag{5}$$

- the numerator is a measure of error
- the denominator is a measure of the standard deviation
- when $|r_i| \geq 2$ of $|r_i| \geq 3$ we will flag the point as an anomaly

Because we want to know how much the parameters vary if we remove a single data point from the data set we have the following:

$$\beta^\star - (\beta^{(-i)})^\star = \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i(y_i - \mathbf{x}_i^T\beta)}{1 - h_{ii}} \tag{6}$$

Proof. Homework.

Fin.