Data - Questions

Thursday 24th November, 2016

1 Introduction

2 Questions About the Data

3 Linear Regression

4 Conclusion

Order of the Day

- what type of questions can be asked (and answered) regarding the data
  - a primer for next topics in statistical inference
- introduction to two important and basic machine learning tasks

1 Introduction

2 Questions About the Data

3 Linear Regression

4 Conclusion

Questions About the Data

Israeltech challenge

	name	filming time $(years)$	budget $(10^6\$)$	profit $(10^6\$)$	genre
	Avatar	1.5	350	650	action
	Titanic	0.8	300	500	drama
V	Die Hard	0.5	-	350	-
V	Looper	0.6	-	400	-
V	Fight Club	0.4	-	700	-
	Inception	0.7	250	400	action
	⋮	⋮	⋮	⋮	⋮

what can we ask about it?

discussion

Questions About the Data

- how was it generated?
- was it generated the same way as another data set D'?
- is it surprising?
- are data elements A and B dependent on each other?
- which category do the points in V belong to?
- what are the missing values for data element A?

Categories and Missing Values

The last two questions:

- which category do the points in V belong to?
- what are the missing values for data element A?

are closely related

Categories and Missing Values

The last two questions:

- which category do the points in V belong to?
  - estimate a missing $\boxed{categorical}$ value
  - example: $genre$
  - task: **classification**

- what are the missing values for data element A?
  - estimate a missing $\boxed{numerical}$ value
  - example: $budget$
  - task: **regression**

1 Introduction

2 Questions About the Data

3 Linear Regression

4 Conclusion

Linear Regression

Data

X^1	X^2	$\cdots$	X^p	Y
1	200	$\cdots$	-0.05	5
1.01	400	$\cdots$	-0.06	7
		$\vdots$		$\vdots$
1.1	460	$\cdots$	-0.1	-
1.5	430	$\cdots$	-0.08	-

Task

- assume $\forall i = 1, \ldots, n : Y_i \approx \sum_{j=1}^{p} \beta^j X_i^j$
- estimate $\beta^1, \ldots, \beta^p$

Linear Regression

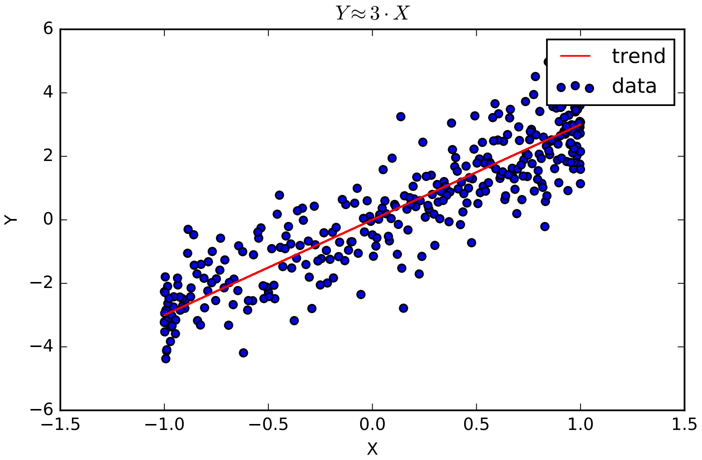
Task

- assume $\forall i = 1, \ldots, n : Y_i \approx \sum_{j=1}^{p} \beta^j X_i^j$
- estimate $\beta^1, \ldots, \beta^p$

Naming

notation	name
$X^1, \ldots, X^p$	independent/explanatory variables
Y	target/dependent/explained
$\epsilon_i = Y_i - \sum_{j=1}^{p} \beta^j X_i^j$	residuals
$\beta^1, \ldots, \beta^p$	coefficients

Example: Trend Line



Estimating the Coefficients

Israeltech challenge

First we transform our equations to matrix notation

$$\forall i = 1, \ldots, n : Y_i \approx \sum_{j=1}^{p} \beta^j X_i^j$$

system of linear equations

Estimating the Coefficients

First we transform our equations to matrix notation

$$\begin{pmatrix} \beta^1 \cdot X_1^1 + \beta^2 \cdot X_1^2 + \ldots + \beta^p \cdot X_1^p \\ \vdots \\ \beta^1 \cdot X_n^1 + \beta^2 \cdot X_n^2 + \ldots + \beta^p \cdot X_n^p \end{pmatrix} \approx \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

Estimating the Coefficients

First we transform our equations to matrix notation

$$\begin{pmatrix} \beta^1 \cdot X_1^1 & \beta^2 \cdot X_1^2 & \dots & \beta^p \cdot X_1^p \\ & \vdots & & \\ \beta^1 \cdot X_n^1 & \beta^2 \cdot X_n^2 & \dots & \beta^p \cdot X_n^p \end{pmatrix} \cdot \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \approx \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

Estimating the Coefficients

First we transform our equations to matrix notation

$$\begin{pmatrix} X_1^1 & X_1^2 & \dots & X_1^p \\ & \vdots & & \\ X_n^1 & X_n^2 & \dots & X_n^p \end{pmatrix} \cdot \begin{pmatrix} \beta^1 \\ \vdots \\ \beta^p \end{pmatrix} \approx \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

in short:

$$X \cdot \beta \approx Y$$

Estimating the Coefficients

Let's solve:

$$X \cdot \beta = Y$$

problem:

X	Y
2	6
3	9
10	30
1	3.001

discuss

Estimating the Coefficients

Idea

solving

$$X \cdot \beta = Y$$

is equivalent to demanding

$$X \cdot \beta - Y = 0$$

instead we can minimize

$$\arg\min_{\beta} \, |X \cdot \beta - Y|$$

problem - non differentiable

Least Squares

Let's solve:

$$\arg\min_{\beta} \|X \cdot \beta - Y\|^2$$

- differentiable
- convex (unique minimizer)

solution:

$$\boxed{\beta = \left(X^T \cdot X\right)^{-1} \cdot X^T \cdot Y}$$

Linearity

So we can solve:

- $Y \approx \beta^1 \cdot 3X^1$
- $Y \approx \beta^1 \cdot 3X^1 + \beta^2 \cdot 5X^2$

But what about?

- $Y \approx \beta^1 \cdot 3X^1 - \beta^2 \cdot 2(X^1)^2$
- $Y \approx \beta^1 \cdot 3X^1 + \beta^0 \cdot 1$
- $Y \approx \beta^1 \cdot 3\sin\left(X^1\right) + \beta^2 \cdot 5\cos\left(X^2\right)$

Linearity

We can simply transform:

X^1	X^2
2	6
3	9
$\vdots$	$\vdots$

$\longrightarrow$

$3X^1$	$-2(X^1)^2$
2	-8
3	-18
$\vdots$	$\vdots$

Linearity

We can simply transform:

X^1	X^2
2	6
3	9
$\vdots$	$\vdots$

$\longrightarrow$

$3X^1$	1
2	1
3	1
$\vdots$	$\vdots$

Linearity

We can simply transform:

X^1	X^2
2	6
3	9
$\vdots$	$\vdots$

$\longrightarrow$

$3\sin\left(X^1\right)$	$5\cos\left(X^1\right)^2$
2.73	0.42
-4.80	4.56
$\vdots$	$\vdots$

Linearity

- the linearity is in the regression coefficients
- strictly speaking
  - polynomial regression
  - ...
- in practice - same solution

Linearity

what can't we solve?

- $Y \approx \sin\left(\beta^1 \cdot X^1\right)$
- $Y \approx \beta\left(X^1\right) \cdot X^1$
- ...

Intercept

X^1	X^2
2	6
3	9
$\vdots$	$\vdots$

$\longrightarrow$

$3X^1$	1
2	1
3	1
$\vdots$	$\vdots$

in this case the result of our regression would be coefficients β^0, β^1 that satisfy

$$\beta^1 \cdot 3X^1 + \beta^0 \cdot 1 \approx Y$$

β^0 the coefficient for the constant term is named the intercept

Assumptions

- $X^1, \ldots, X^p$ - linearly independent
- more samples than variables (overdetermined)
- low measurement error $X^1, \ldots, X^p$
- fixed variance - homoscedasticity
- $\epsilon_1, \ldots, \epsilon_n$ - statistically independent

there are more

Another Direction

- least squares - cool trick
- what about some statistics?

Another Direction

Israeltech challenge

Residuals (1D)

$$\forall i = 1, \ldots, n : \epsilon_i = Y_i - X_i \cdot \beta$$

- $\epsilon_1, \ldots, \epsilon_n$ are independent
- we assume $E[\epsilon_i] = 0$ (always) explain
- let's assume that $\forall i = 1, \ldots, n : \epsilon_i \sim \mathcal{N}\left(0, \sigma_\epsilon^2\right)$

- $Y_i = X_i \cdot \beta + \epsilon_i$

Another Direction

Israeltech challenge

Residuals (1D)

$$\forall i = 1, \ldots, n : \epsilon_i = Y_i - X_i \cdot \beta$$

- $\epsilon_1, \ldots, \epsilon_n$ are independent
- we assume $E\left[\epsilon_i\right] = 0$ (always) explain
- let's assume that $\forall i = 1, \ldots, n : \epsilon_i \sim \mathcal{N}\left(0, \sigma_\epsilon^2\right)$

- $(Y_i \mid X_i, \beta) \sim \mathcal{N}\left(X_i \cdot \beta, \sigma_\epsilon^2\right)$ independently

notation, what now?

Maximum Likelihood Estimation

$$P(Y \mid X, \beta) = \prod_{i=1}^{n} P(Y_i \mid X_i, \beta)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} e^{-\frac{(Y_i - X_i \cdot \beta)^2}{2\sigma_\epsilon^2}}$$

$$= \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} e^{-\frac{1}{2\sigma_\epsilon^2} \sum_{i=1}^{n} (Y_i - X_i \cdot \beta)^2}$$

Maximum Likelihood Estimation

Israeltech challenge

$$
\begin{aligned}
\beta_{mle} &= \arg\max_{\beta} P\left(Y \mid X, \beta\right) \\
&= \arg\max_{\beta} \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} e^{-\frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{n}\left(Y_i - X_i \cdot \beta\right)^2} \\
&= \arg\max_{\beta} \; -\frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{n}\left(Y_i - X_i \cdot \beta\right)^2 \\
&= \arg\min_{\beta} \; \sum_{i=1}^{n}\left(Y_i - X_i \cdot \beta\right)^2
\end{aligned}
$$

Maximum Likelihood Estimation

$$\boxed{\beta_{mle} = \arg\min_{\beta} \|Y - X \cdot \beta\|^2}$$

Metrics

Israeltech challenge

how do we measure success?

- Mean Squared Error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i^{true} - Y_i\right)^2$$

- Coefficient of Determination r^2

$$r^2 = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - X_i \cdot \beta\right)^2}{\sum_{i=1}^{n}\left(Y_i - \mathbb{E}\left[Y_i\right]\right)^2}$$

discuss

1 Introduction

2 Questions About the Data

3 Linear Regression

4 Conclusion

Questions About the Data

- how was it generated?
- was it generated the same way as another data set D'? - Statistical tests
- is it surprising? - hypothesis testing
- are data elements A and B dependent on each other? - Statisical tests
- which category do the points in V belong to? - Data Science
- what are the missing values for data element A? - Linear Regression