GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

LECTURE 6: REGRESSION & REGULARIZATION

LAST TIME:

I. WHAT IS MACHINE LEARNING?
II. MACHINE LEARNING PROBLEMS
III. CLASSIFICATION PROBLEMS
IV. KNN CLASSIFICATION (& EXERCISES)

QUESTIONS?

5. INTRO TO MACHINE LEARNING & KNN (LAST WEEK)
6. LINEAR REGRESSION & LINEAR ALGEBRA (TODAY)
7. REGRESSION & REGULARIZATION
8. STATISTICS & BAYES
9. DECISION TREES
10. RECAP SUPERVISED LEARNING

0. PRESENTATIONS ASSIGNMENT DATA EXPLORATION

I. LINEAR ALGEBRA & NUMPY
II. INTRO TO REGRESSION
III. MATH BEHIND THE SCENES (OPTIONAL)

I. LINEAR ALGEBRA

II. LINEAR REGRESSION

	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

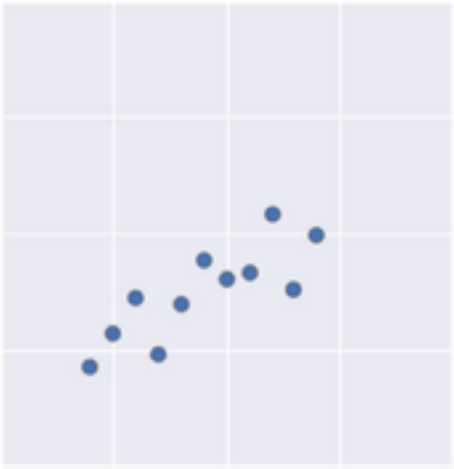
Q: What is a **regression** *model?*

Q: What is a **regression** *model?*

A: A functional relationship between input & response variables

Q: What is a **regression** *model?*

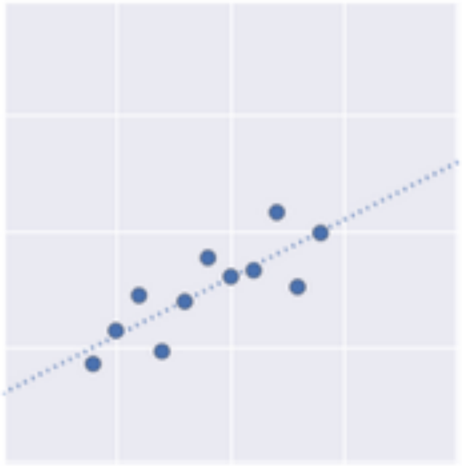
A: A functional relationship between input & response variables



Source: Anscombe's Quartet

Q: What is a **regression** *model?*

A: A functional relationship between input & response variables



Source: Anscombe's Quartet

Q: What is a **regression** *model?*

A: A functional relationship between input & response variables

The **simple linear regression** *model captures a linear relationship between a single input variable x and a response variable y:*

Q: What is a **regression** *model?*

A: A functional relationship between input & response variables

The **simple linear regression** *model captures a linear relationship between a single input variable x and a response variable y:*

$$y = \alpha + \beta x + \varepsilon$$

Q: What do the terms in this model mean?

$$y = \alpha + \beta x + \varepsilon$$

Q: What do the terms in this model mean?

$$y = \alpha + \beta x + \varepsilon$$

y = **response variable** *(the one we want to predict)*

Q: What do the terms in this model mean?

$$y = \alpha + \beta x + \varepsilon$$

y = **response variable** *(the one we want to predict)*

x = **input variable** *(the one we use to train the model)*

Q: What do the terms in this model mean?

$$y = \alpha + \beta x + \varepsilon$$

y = **response variable** *(the one we want to predict)*

x = **input variable** *(the one we use to train the model)*

α = **intercept** *(where the line crosses the y-axis)*

Q: What do the terms in this model mean?

$$y = \alpha + \beta x + \varepsilon$$

y = **response variable** *(the one we want to predict)*

x = **input variable** *(the one we use to train the model)*

α = **intercept** *(where the line crosses the y-axis)*

β = **regression coefficient** *(the model "parameter")*

Q: What do the terms in this model mean?

$$y = \alpha + \beta x + \varepsilon$$

y = **response variable** *(the one we want to predict)*

x = **input variable** *(the one we use to train the model)*

α = **intercept** *(where the line crosses the y-axis)*

β = **regression coefficient** *(the model "parameter")*

ε = **residual** *(the prediction error)*

$$y = \alpha + \beta x + \varepsilon$$

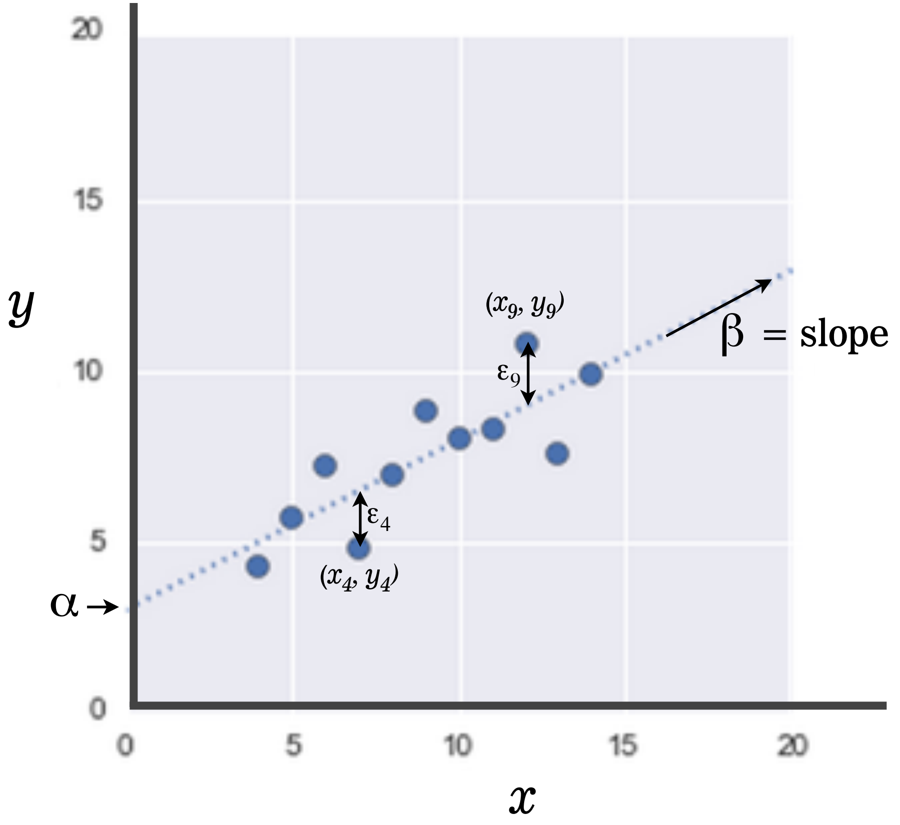
y = **response variable**

x = **input variable**

α = **intercept**

β = **regression coefficient**

ε = **residual** *(the error)*



Source: Anscombe's Quartet

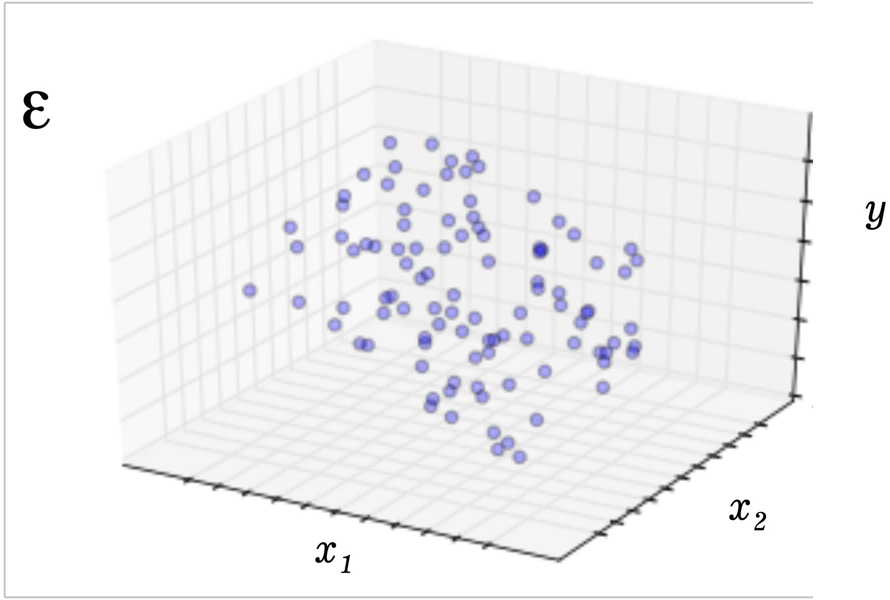
We can extend this model to several input variables, giving us the **multiple linear regression** *model:*

We can extend this model to several input variables, giving us the **multiple linear regression** *model:*

$$y = \alpha + \beta_1 x_1 + \ldots + \beta_n x_n + \varepsilon$$

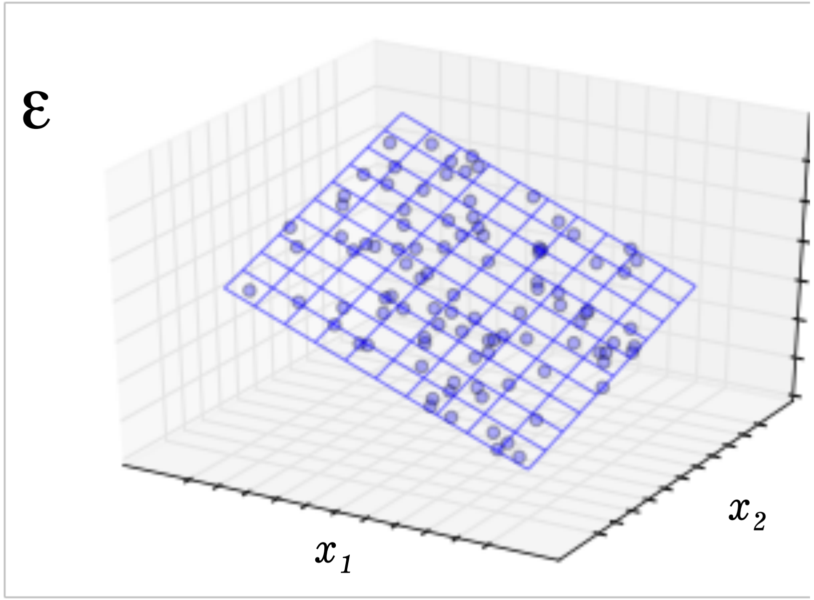
We can extend this model to several input variables, giving us the **multiple linear regression** *model:*

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$



We can extend this model to several input variables, giving us the **multiple linear regression** *model:*

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$



Q: How do we fit a regression model to a dataset?

Q: How do we fit a regression model to a dataset?

A: In theory, minimize the sum of the squared residuals (OLS).

Q: How do we fit a regression model to a dataset?

A: In theory, minimize the sum of the squared residuals (OLS).

In practice, any respectable piece of software will do this for you.

Q: How do we fit a regression model to a dataset?

A: In theory, minimize the sum of the squared residuals (OLS).

In practice, any respectable piece of software will do this for you.

But again, if you get serious about regression, you should learn how this works!

QUESTION

HOW
DO YOU
MEASURE
THE
QUALITY?

supervised	*making predictions*
unsupervised	*extracting structure*

supervised	*test out your predictions*

	continuous	*categorical*
supervised	*regression*	*classification*

R^2
mean absolute error
mean squared error

Accuracy
(% correct predictions)
and other metrics

III: MATH BEHIND REGRESSION

Q: How do we fit a regression model to a dataset?

Given the **multiple linear regression** *model:*

$$y = \alpha + \beta_1 x_1 + \ldots + \beta_n x_n + \varepsilon$$

Given the **multiple linear regression** *model:*

$$y = \alpha + \beta_1 x_1 + \ldots + \beta_n x_n + \varepsilon$$

$$\hat{y} = \alpha + \beta_1 x_1 + \ldots + \beta_n x_n$$

The residual (error) is equal to the observed y minus the predicted ŷ

$$\varepsilon = -y + \alpha + \beta_1 x_1 + \ldots + \beta_n x_n$$

$$= \hat{y} - y$$

*Define a **cost function** J of the parameters α and βs*

$$J(\alpha,\beta) = \Sigma\,(-y + \alpha + \beta_1 x_1 + \ldots + \beta_n x_n)^2$$

which sums the squares of all prediction errors

Then we're looking for those β s where J has its minimum:

$$\min_{\alpha,\beta} J = \min_{\alpha,\beta} \sum (-y + \alpha + \beta_1 x_1 + \ldots + \beta_n x_n)^2$$

*Define a **cost function** J of the parameters β*

$$\min_{\alpha,\beta} J = \min_{\alpha,\beta} \Sigma \left(-y + \alpha + \beta_1 x_1 + \ldots + \beta_n x_n\right)^2$$

and find where J has its minimum

This is called the Ordinary Least Squares (OLS) method

Let's simplify notation using linear algebra

Given the **multiple linear regression** *model:*

$$y = \alpha + \beta_1 x_1 + \ldots + \beta_n x_n + \varepsilon$$

Given the **multiple linear regression** *model:*

$$\text{N samples} \left\{ \begin{array}{l} y_1 = \alpha + \beta_1 x_{11} + \ldots + \beta_n x_{1n} + \varepsilon_1 \\ y_2 = \alpha + \beta_1 x_{21} + \ldots + \beta_n x_{2n} + \varepsilon_2 \\ y_3 = \alpha + \beta_1 x_{21} + \ldots + \beta_n x_{3n} + \varepsilon_3 \\ y_4 = \alpha + \beta_1 x_{31} + \ldots + \beta_n x_{4n} + \varepsilon_4 \end{array} \right.$$

Given the **multiple linear regression** *model:*

$$\textit{N samples}\left\{\begin{array}{l} y_1 = \alpha + \beta_1 x_{11} + \ldots + \beta_n x_{1n} + \varepsilon_1 \\ y_2 = \alpha + \beta_1 x_{21} + \ldots + \beta_n x_{2n} + \varepsilon_2 \\ y_3 = \alpha + \beta_1 x_{21} + \ldots + \beta_n x_{3n} + \varepsilon_3 \\ \underbrace{y_4}_{\textit{labels}} = \alpha + \underbrace{\beta_1 x_{31} + \ldots + \beta_n x_{4n}}_{\textit{n features}} + \varepsilon_4 \end{array}\right.$$

Given the **multiple linear regression** *model:*

$$
\text{N samples}\left\{
\begin{aligned}
y_1 &= \alpha + \beta_1 x_{11} + \ldots + \beta_n x_{1n} + \varepsilon_1 \\
y_2 &= \alpha + \beta_1 x_{21} + \ldots + \beta_n x_{2n} + \varepsilon_2 \\
y_3 &= \alpha + \beta_1 x_{21} + \ldots + \beta_n x_{3n} + \varepsilon_3 \\
y_4 &= \alpha + \beta_1 x_{31} + \ldots + \beta_n x_{4n} + \varepsilon_4
\end{aligned}
\right.
$$

labels *n features* *residuals (errors)*

We can summarize all samples in vectors

$$\mathbf{y} = \alpha + \beta_1\ \mathbf{x}_1 + \ \dots\ + \ \beta_n \mathbf{x}_n + \boldsymbol{\varepsilon}$$

y, ***x*** *and* **$\boldsymbol{\varepsilon}$** *are vectors of dimension N,*

where N is the number of samples (observations)

And we can summarize all features ...

$$\mathbf{y} = \alpha + \beta_1 \underbrace{\mathbf{x}_1 + \ \dots \ + \ \beta_n \mathbf{x}_n}_{\textit{features}} + \boldsymbol{\varepsilon}$$

*And we can summarize all features in a matrix **X***

$$\mathbf{y} = \alpha + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

X *is an N×n-matrix,*
with N rows for each observation (samples)
and n columns for each feature

β *is a n-dimensional vector,*
with coefficients for each feature

So we can write the linear regression, using linear algebra, as

$$\mathbf{y} = \alpha + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

So we can write the linear regression, using linear algebra, as

$$\mathbf{y} = \alpha + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

y = **response variable** *(the one we want to predict)* — *N-dim vector*

X = **input variable** *(the one we use to train the model)* — *N×n-matrix*

α = **intercept** *(where the line crosses the y-axis)* — *scalar*

β = **regression coefficient** *(the model "parameter")* — *n-dim vector*

ε = **residual** *(the prediction error)* — *N-dim vector*

So we can write the linear regression, using linear algebra, as

$$\mathbf{y} = \alpha + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

And hence the cost function as the norm of the residual vector $\boldsymbol{\varepsilon}$

$$J(\alpha, \boldsymbol{\beta}) = \left| -\mathbf{y} + \alpha + \mathbf{X}\boldsymbol{\beta} \right|$$

So we can write the linear regression, using linear algebra, as

$$\mathbf{y} = \alpha + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

And hence the cost function as the norm of the residual vector $\boldsymbol{\varepsilon}$

$$J(\alpha, \boldsymbol{\beta}) = \left| -\mathbf{y} + \alpha + \mathbf{X}\boldsymbol{\beta} \right|$$

$$= \sum_{i=1}^{N} \left(-y_i + \alpha + \mathbf{x}_i\boldsymbol{\beta} \right)^2$$

So we can write the linear regression, using linear algebra, as

$$\mathbf{y} = \alpha + \mathbf{X\beta} + \boldsymbol{\varepsilon}$$

And hence the cost function as the norm of the residual vector $\boldsymbol{\varepsilon}$

$$J(\alpha, \boldsymbol{\beta}) = \left| -\mathbf{y} + \alpha + \mathbf{X\beta} \right|$$

which is a function $\mathbb{R}^{n+1} \to \mathbb{R}$, *for which we want to know the minimum*

Often an artificial 0th column of 1s is added to $\mathbf{X}$*, such that we get*

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

and $\alpha = \beta_0$*, which is often easier in notation and computations*

$$J(\boldsymbol{\beta}) = |-\mathbf{y} + \mathbf{X}\boldsymbol{\beta}|$$

$\mathbf{X}$ is now an $N \times (n+1)$-matrix
$\mathbf{X} = (1, \mathbf{x}_1, \mathbf{x}_2, \ldots)$

which is a function $\mathbb{R}^{n+1} \to \mathbb{R}$*, for which we want to know the minimum*

The OLS method has a closed-form solution for the parameter β

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

The OLS method has a closed-form solution for the parameter β

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

*Generally, though, you do **not** have a closed-form solution to find the minimum of the cost function. In that case, the **gradient descent** algorithm will help you out. (To Be Continued.)*

QUESTION

HOW DO YOU MEASURE THE QUALITY?

An intuitive measure is the ***mean absolute error***

$$\mathbf{MAE} = \frac{1}{N}\Sigma\,|\varepsilon| \quad = \frac{1}{N}\Sigma\,|\,y - \hat{y}\,|$$

*An intuitive measure is the **mean absolute error***

$$\mathbf{MAE} = \frac{1}{N}\Sigma\,|\varepsilon| \quad = \frac{1}{N}\Sigma\,|\,y - \hat{y}\,|$$

*Another widely used one is the **mean squared error***

$$\mathbf{MSE} = \frac{1}{N}\Sigma\,\varepsilon^2 \quad = \frac{1}{N}\Sigma\,(y - \hat{y})^2$$

An intuitive measure is the ***mean absolute error***

$$\mathbf{MAE} = \frac{1}{N}\Sigma\ |\varepsilon| \quad = \frac{1}{N}\Sigma\ |\ y - \hat{y}\ |$$

Another widely used one is the ***mean squared error***

$$\mathbf{MSE} = \frac{1}{N}\Sigma\ \varepsilon^2 \quad = \frac{1}{N}\Sigma\ (y - \hat{y})^2$$

Note that the OLS method minimizes the MSE

The most commonly used measure is R^2, which is a ratio of variance

$$\mathbf{MSE} = \frac{1}{N}\Sigma\ (y - \widehat{y})^2$$

mean squared error

The most commonly used measure is R^2, which is a ratio of variance

$$\mathbf{MSE} = \frac{1}{N}\Sigma\,(y - \hat{y})^2$$ *mean squared error*

$$\mathrm{Var}\,\mathbf{y} = \frac{1}{N}\Sigma\,(y - \overline{y})^2$$ *variance of observed data*

The most commonly used measure is R^2, which is a ratio of variance

$$R^2 = 1 - \frac{\frac{1}{N}\sum (y - \hat{y})^2}{\frac{1}{N}\sum (y - \overline{y})^2}$$

mean squared error

variance of observed data

The most commonly used measure is R^2, which is a ratio of variance

$$R^2 = 1 - \frac{\Sigma\,(y - \hat{y})^2}{\Sigma\,(y - \overline{y})^2}$$

The most commonly used measure is R^2, which is a ratio of variance

$$R^2 = 1 - \frac{\Sigma\,(y - \hat{y})^2}{\Sigma\,(y - \overline{y})^2}$$

$R^2 = 1$ iff all errors are zero

The most commonly used measure is R^2, which is a ratio of variance

$$R^2 = 1 - \frac{\Sigma\,(y - \hat{y})^2}{\Sigma\,(y - \overline{y})^2}$$

$R^2 = 1$ iff all errors are zero

$R^2 < 0$ is possible (very bad)

DISCUSSION