

Introduction to Machine Learning (and Notation)

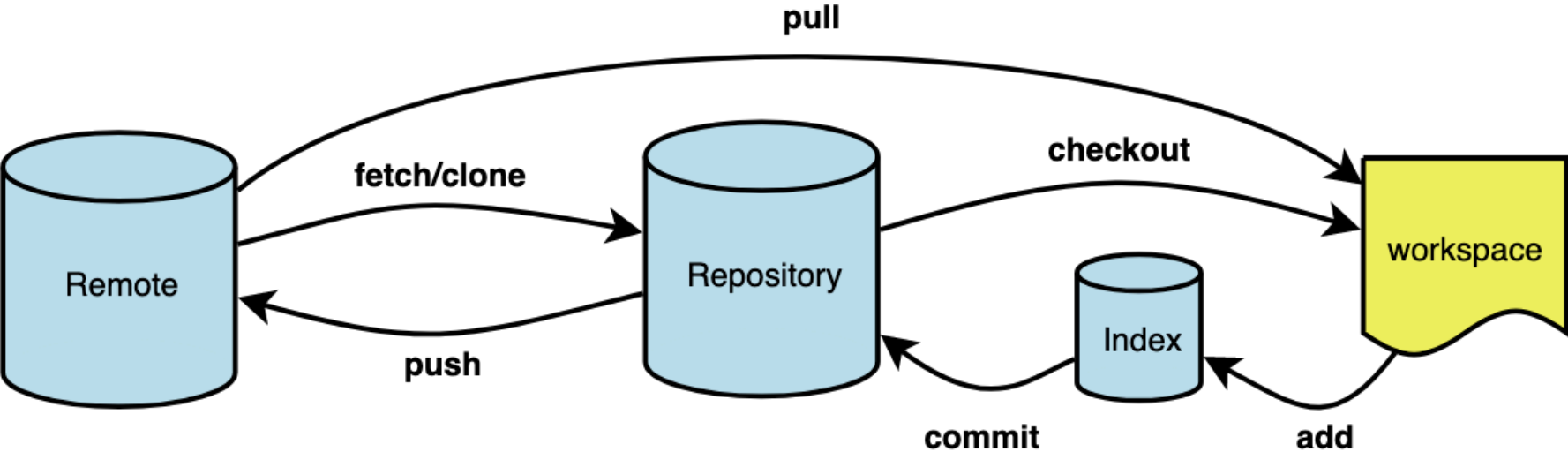
David I. Inouye

Wednesday, August 28, 2019

Announcements

- ▶ EE 134 is Liming's temporary location for office hours
- ▶ Make sure you are working on selecting your 3 research papers
- ▶ Homework 1 is posted
 - ▶ <https://www.davidinouye.com/course/ece57000-fall-2019/hw1/>
 - ▶ See link on Piazza to initialize git repository

GIT Structure



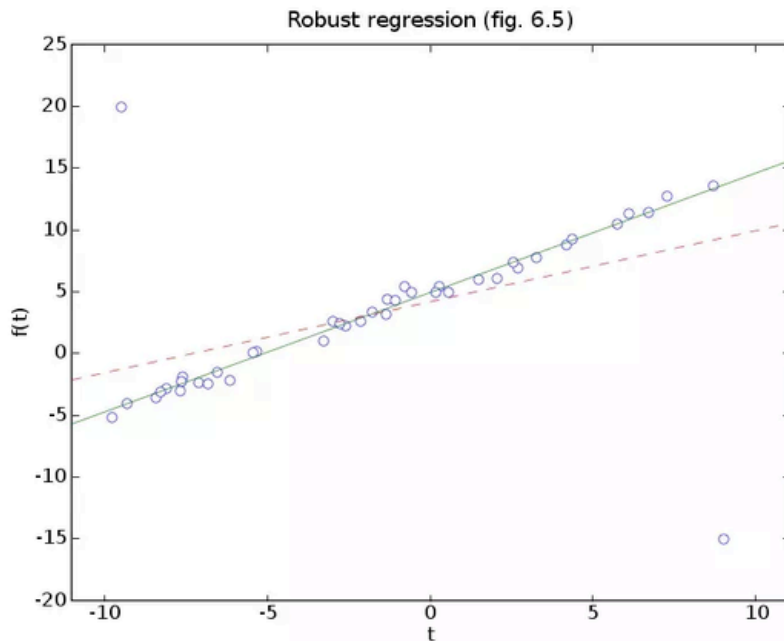
<https://illustrated-git.readthedocs.io/en/latest/>

Correlation analysis vs linear regression analysis (ordinary least squares, OLS)

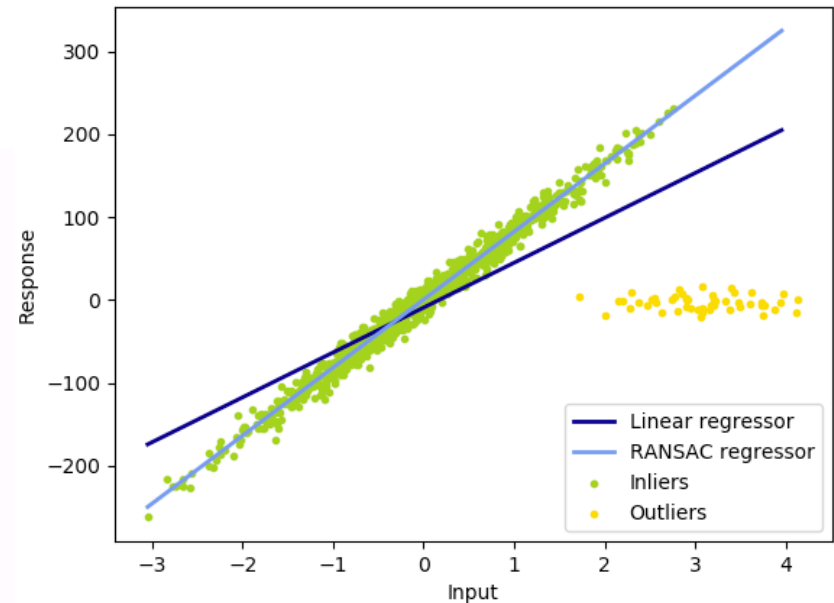
- ▶ The goals are different
 - ▶ Correlation – Measure linear dependence
 - ▶ Linear regression – Predict y given x
- ▶ The output of the two methods is different
 - ▶ Correlation – Output is **single number** ρ between -1 and 1 that measures **linear** dependency between variables
 - ▶ Linear regression – Outputs **function** $y \approx f(x)$, which has a linear form, i.e. $f(x) = bx + a$

Correlation analysis vs linear regression analysis (ordinary least squares, OLS)

- ▶ We must assume we are doing ordinary least squares (OLS) linear regression



<https://www.quora.com/How-is-Robust-Regression-different-from-standard-OLS>



https://scikit-learn.org/stable/auto_examples/linear_model/plot_ransac.html

Correlation analysis vs linear regression analysis (ordinary least squares, OLS)

▶ However, the output of correlation analysis and the parameters of linear regression have some relationship

▶ Let s_x be the standard deviation of x and similarly for y , let $s_{x,y}$ be the covariance between x and y

▶ We have this simple relationship $p = b \left(\frac{s_x}{s_y} \right)$ or

similarly $b = p \left(\frac{s_y}{s_x} \right)$

▶ We can derive this from the following:

▶ $p = \frac{s_{x,y}}{s_x s_y}$, correlation formula

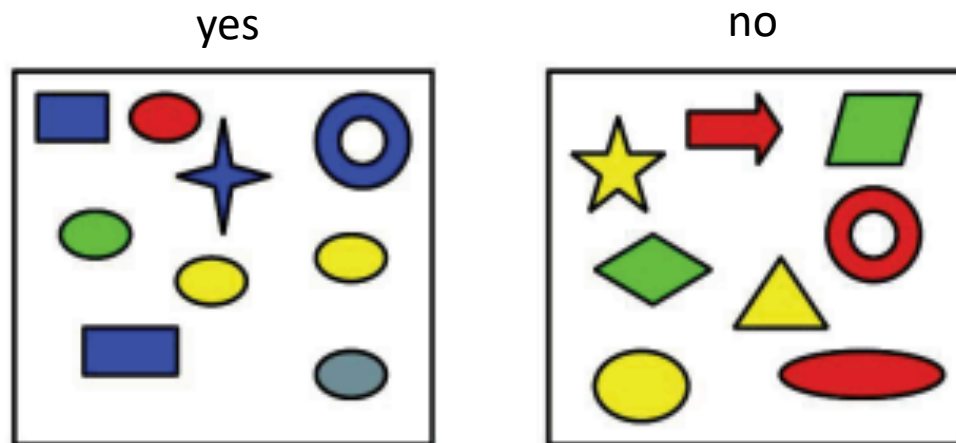
▶ $b = \frac{s_{x,y}}{s_x^2}$, $a = \bar{y} - b\bar{x}$, linear regression formula

<https://www.stat.berkeley.edu/~rabbee/correlation.pdf>

Correlation analysis vs linear regression analysis (ordinary least squares, OLS)

- ▶ Cannot derive one from the other directly.
Must know standard deviations.
- ▶ Correlation is invariant to scaling of x and y ,
both good and bad
- ▶ The output of linear regression does not retain
any information about the original distribution
(it's just a line)

The dataset cannot determine the task, rather **the context** determines the task



d features/attributes/covariates

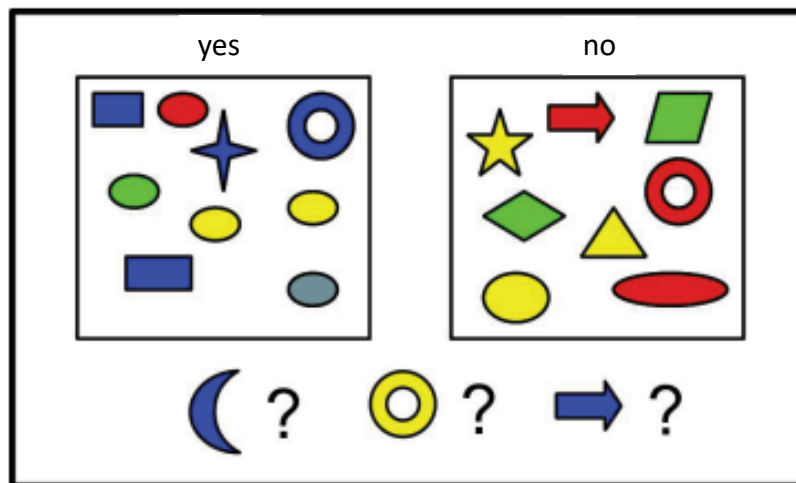
| Color | Shape | Size (cm) | Is it good? |
|-------|---------|-----------|-------------|
| Blue | Square | 10 | yes |
| Red | Ellipse | 2.4 | yes |
| Red | Ellipse | 20.7 | no |

n samples/
observations/
examples

Dataset: Age and top running/walking speed
What is the task and what are x and y ?

- ▶ Suppose you are a running shoe company; you would like to make personalized products for each person but you can only create three product lines given this data
- ▶ Suppose you are a policeman and a suspect outran you; you would like to guess more about the suspect
- ▶ Suppose you are a criminal and you know the one policeman on duty

Generalization *beyond* the training set is the main goal of learning

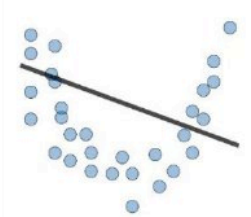

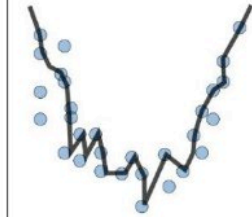
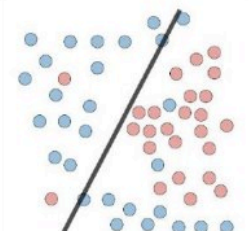
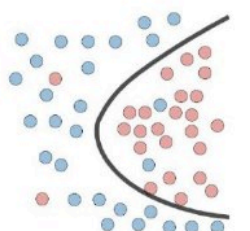
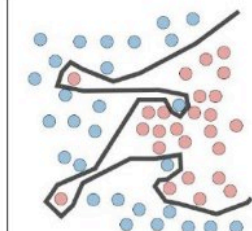
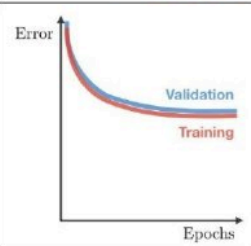
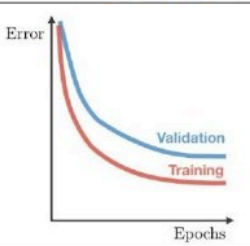
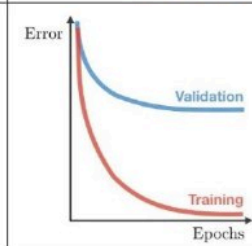


d features/attributes/covariates

| | | Color | Shape | Size (cm) | Is it good? | | |
|---|-------|-------|---------|-----------|-------------|-------|--|
| n samples/ observations/ examples | x_1 | Blue | Square | 10 | yes | y_1 | |
| | x_2 | Red | Ellipse | 2.4 | yes | y_2 | |
| | | Red | Ellipse | 20.7 | no | | |

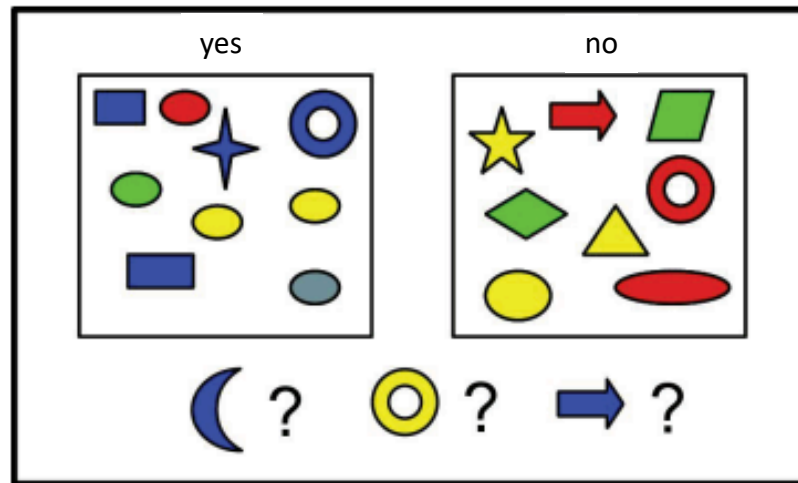
Example from Machine Learning: A Probabilistic Perspective, Ch. 1, Kevin P. Murphy, 2012.

Generalization *beyond* the training set is the main goal of learning

| | Underfitting | Just right | Overfitting |
|----------------|--|---|--|
| Symptoms | <ul style="list-style-type: none"> - High training error - Training error close to test error - High bias | <ul style="list-style-type: none"> - Training error slightly lower than test error | <ul style="list-style-type: none"> - Low training error - Training error much lower than test error - High variance |
| Regression |  |  |  |
| Classification |  |  |  |
| Deep learning |  |  |  |
| Remedies | <ul style="list-style-type: none"> - Complexify model - Add more features - Train longer | | <ul style="list-style-type: none"> - Regularize - Get more data |

Original source for figure unknown.

Probability can *formalize* the handling of ambiguity



d features/attributes/covariates

| | | Color | Shape | Size (cm) | Is it good? | | |
|---|-------|-------|---------|-----------|-------------|-------|--|
| n samples/ observations/ examples | x_1 | Blue | Square | 10 | yes | y_1 | |
| | x_2 | Red | Ellipse | 2.4 | yes | y_2 | |
| | | Red | Ellipse | 20.7 | no | | |

Example from Machine Learning: A Probabilistic Perspective, Ch. 1, Kevin P. Murphy, 2012.

The curse of dimensionality is *unintuitive*

▶ Ratio between unit hypersphere to unit hypercube

▶ 1D : $2/2 = 1$

▶ 2D : $\frac{\pi}{4} = 0.7854$

▶ 3D : $\frac{3\sqrt{\pi}}{8} = 0.5238$

▶ d-dimensions: $V_d(r) = \frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2}+1)} r^d$

▶ Thus, for 10-D: $2.55/2^{10} = 2.55/1024 = 0.00249$

