Comparing Performance Metrics

a.k.a. reproducibility, random seeds and how to not trick yourself

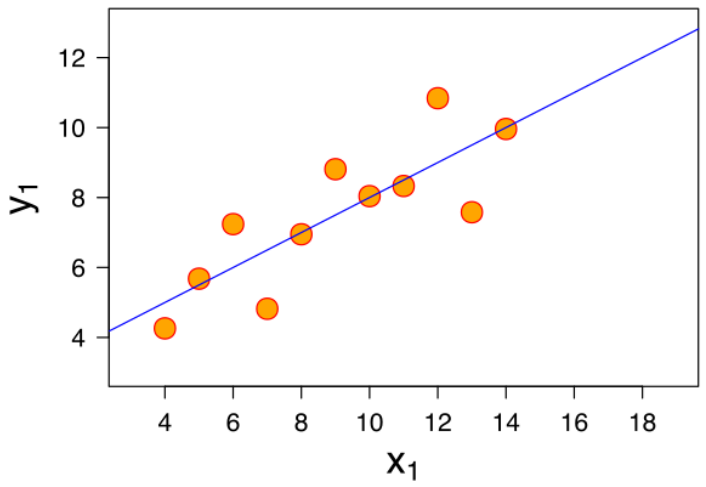
Zoetis Internal Presentation

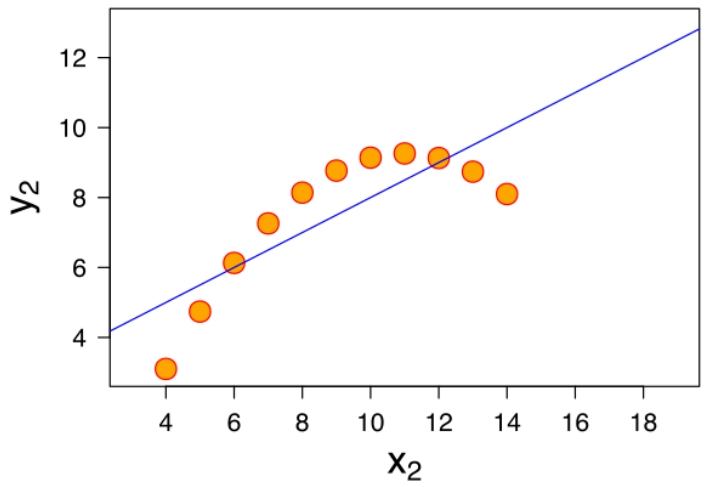
Reading Group
Jan. 2022

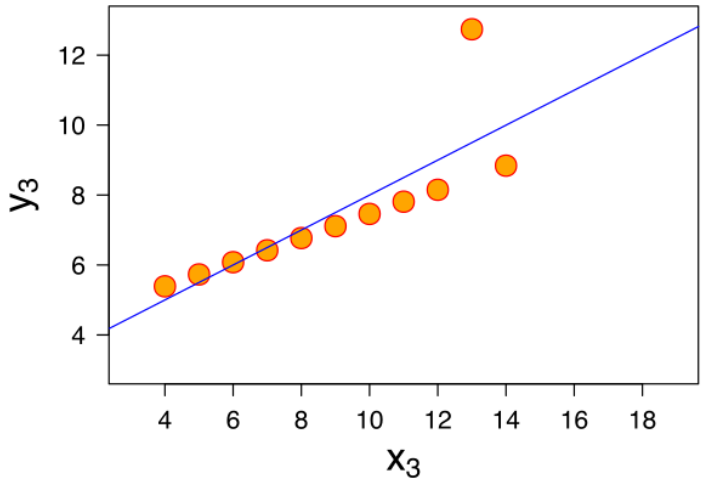
Outline

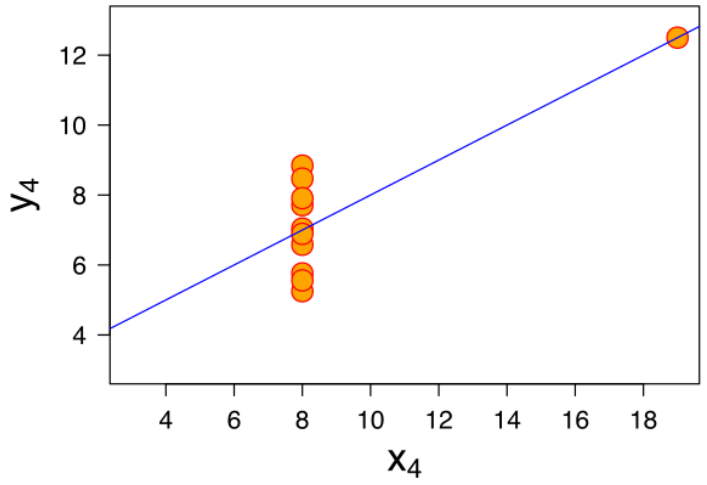
Outline

An Analogy


12
10
8
6
4
y1
4
6
8
10
12
14
16
18
x1


12
10
8
6
4
y2
4
6
8
10
12
14
16
18
x2


12
10
8
6
4
y3
4
6
8
10
12
14
16
18
x3


12
10
8
6
4
y4
4
6
8
10
12
14
16
18
x4

Setting Up the Problem

- Let's assume we are performing a binary classification task
- The data has 50 features and 1,000 labeled examples.
- There are 3 data scientists solving the task in parallel, *each working on a different APPROACH*
- Laura: selects a deep neural network, Bob: selects gradient boosting (AdaBoost), Coop: selects random forests
- Side note: the above approaches are non-deterministic, i.e. multiple trials may lead to multiple different results

Approaches versus Models

Definition

A *learning approach* describes the setup to solving an optimization problem. This includes selection of the *model class*, the loss function, the selection of optimizer, the hyperparameter selections, . . . , i.e. choices a data scientist can make.

Definition

A *model* reflects the weights learned in the process defined by the *learning approach*. These weights can be discovered in a *deterministic* or *non-deterministic* fashion.

Setting Up the Problem

- At the end of the day, one solution needs to be selected to move forward to pilot
- HOW DO WE CHOOSE THE “BEST” SOLUTION?
- Typically, split the data into train/dev/test folds. Train a model on train fold, tune hyperparameters on dev fold, report findings on test fold
  - Even this is hard, more later
- Compare results on test fold, MAYBE with a significance test (typically not though :( )

Initial Findings

Scientist	Method	F1-score
Coop	RandomForest	0.96153
Laura	NeuralNet	0.93617
Bob	AdaBoost	0.84761

- Who's approach should be deployed to production?
- IMPORTANT FOOTNOTE: "For real-world applications, superiority can mean many distinct things that are not related to accuracy"

Outline

Which Significance Tests?

- Two suggested tests
  - Approximate Randomized Test
  - Bootstrap Test
- Both test the null hypothesis that both models would perform equally on the *population* as a whole
- But remember. . . we have a labelled sample from the population, we split that sample into train/dev/test, that splitting required a random seed, each scientist may have biased those splits, etc.
- Let's assume the three scientists discussed and agreed on a fixed train/dev/test split they are all working from to reduce variance

More Choices for Comparison

- Fixed train/dev/test, fixed sig. test: bootstrap
- Still, multiple choices for comparison
  - Single scores on test
  - Best scores on dev
- Problem: comparisons are happening on train/dev sets, NOT ON THE POPULATION

Notation

- We choose two approaches A and B
- Let Ψ be their performance on a chosen metric (accuracy, recall, F1-score, etc.)
- Each approach can have multiple *models*: $A_1, \ldots, A_n$ and $B_1, \ldots, B_m$
- The *best model* A_* and B_* are the configurations that performed best **on the dev set**
- Main idea: $\Psi_{A_*}^{(test)} > \Psi_{B_*}^{(test)} \not\Longrightarrow$ A is a better approach than B

Demonstrating the Issue

- We choose a single approach A and label the alternative $\hat{A}$
- The only difference between A and $\hat{A}$ is their initial configuration, *all controlled by a random seed*
- If significance testing via bootstrap is a good method, we should detect no significant difference between A and $\hat{A}$
- Let's demonstrate this for the models of Laura, Bob and Coop by training 1,000 models with different random seeds for model initialization
- Compute the number of times there is a significant difference between A and $\hat{A}$ as well as the average difference in F1 score for models with significant differences, call this τ

Interesting Results

Scientist	Method	Threshold (τ)	% significant
Coop	RandomForest	0.0401	0.2746
Laura	NeuralNet	0.0313	0.1928
Bob	AdaBoost	0.0567	0.1988

Train vs. Dev vs. Test

- When we select models A_* and B_* based on performance on dev set only, we are assuming monotonicity on performance on the test set; *this assumption may not hold*
- Let's assume instead there are two approaches with the same performance on the dev set i.e. $\Psi_{A_1}^{(dev)} = \Psi_{A_2}^{(dev)}$
- How much can the difference between performance on the test set vary? i.e. $|\Psi_{A_1}^{(test)} - \Psi_{A_2}^{(test)}|$
- The authors find that this difference, EVEN WITH THE SAME PERFORMANCE ON THE DEV SET, can vary up to 3.68 percentage points
- At most conferences, state-of-the-art (SOTA) approaches often beat the baselines by fractions of percentage points, with no significance score even considered

How Do We Fix This?

- Two approaches: one with normality assumption, one without
- Normality: approach A is superior to approach B if and only if the expected test score for A is larger than the expected test score for B
- $\mathbb{E}[\Psi^{(test)}_{A(train,dev,rnd)}] > \mathbb{E}[\Psi^{(test)}_{B(train,dev,rnd)}]$
- The expected values can be approximated by training multiple models (varying the random seed rnd) and conducting a Welch's t-test
- The t-test assumes that the distribution of scores is normal, which may not hold

Non-normality

- Approach A is superior to approach B iff the probability of A is higher to produce a better working model than B
- $P(\Psi^{(test)}_{A(train,dev,rnd)} \geq \Psi^{(test)}_{B(train,dev,rnd)}) > 0.5$
- Choose a sufficiently large number of models and estimate using the Mann-Whitney U test (independent pairs) or the Wilcoxon signed-rank test (dependent pairs)

Summary

- Single metric performance comparisons can be faulty
- Setting random seeds prevents false findings or your risk of chasing the lottery ticket hypothesis
- Better performance comparison schemas exist, yet few are using them