Model Evaluation

20 Oct 2566

Sarun Gulyanon

Agenda

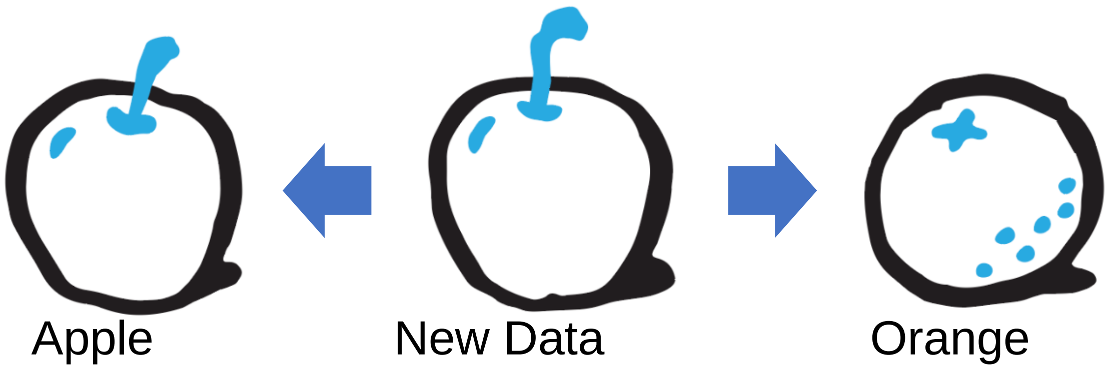
- Model Evaluation
- Case Study : Kaggle

1.

Model Evaluation

Recap: Generalization

- The error rate on new cases is called the *generalization error* (or *out-of-sample error*)
- This value tells you how well your model will perform on instances it has **never seen before**.
- We can only estimate of this error.



https://www.gapminder.org/factfulness/generalization/

Model Evaluation

- Methods for Performance Evaluation
  - How to obtain reliable estimates?

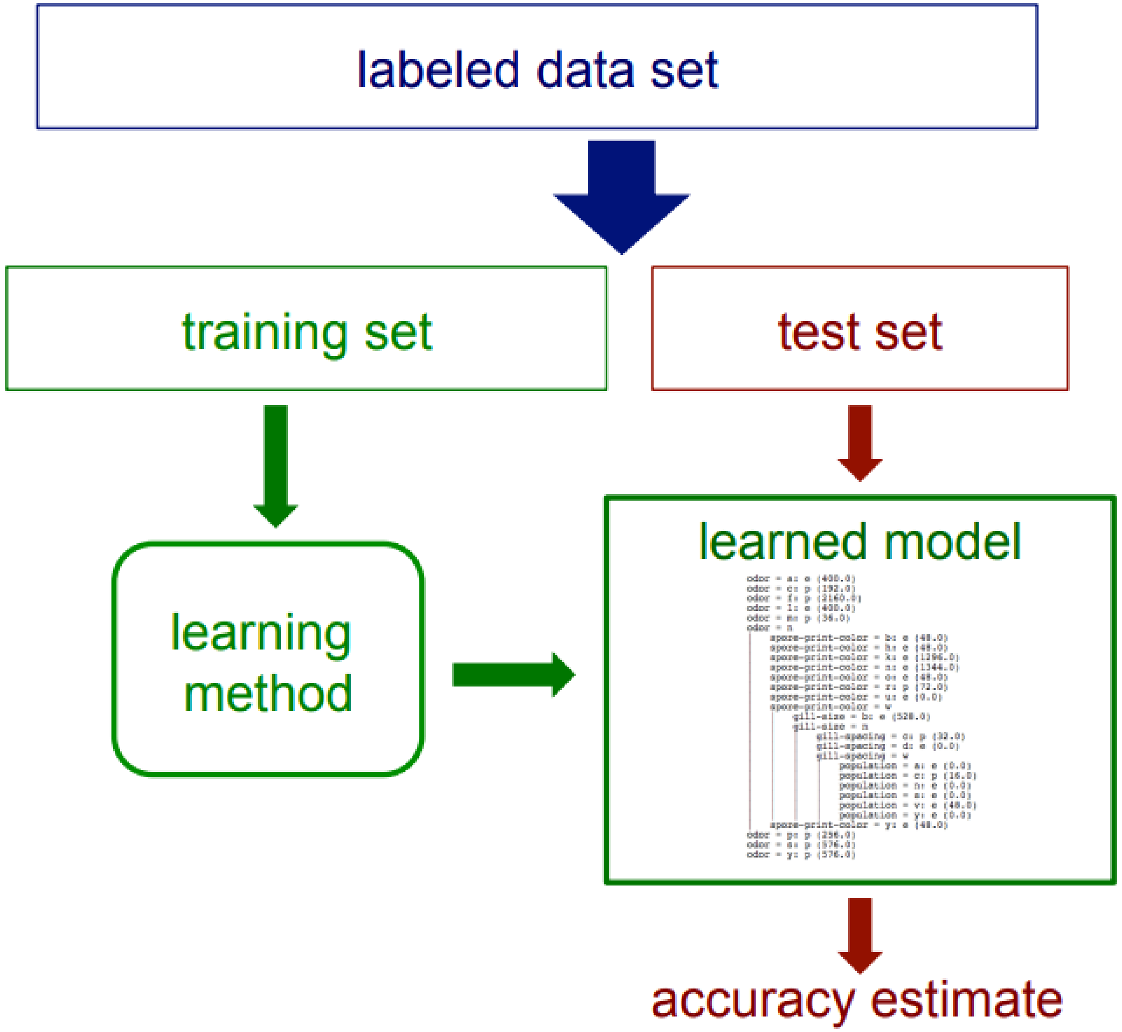
- Metrics for Performance Evaluation
  - How to evaluate the performance of a model?

- Methods for Model Comparison
  - How to compare the relative performance among competing models?

Test Set

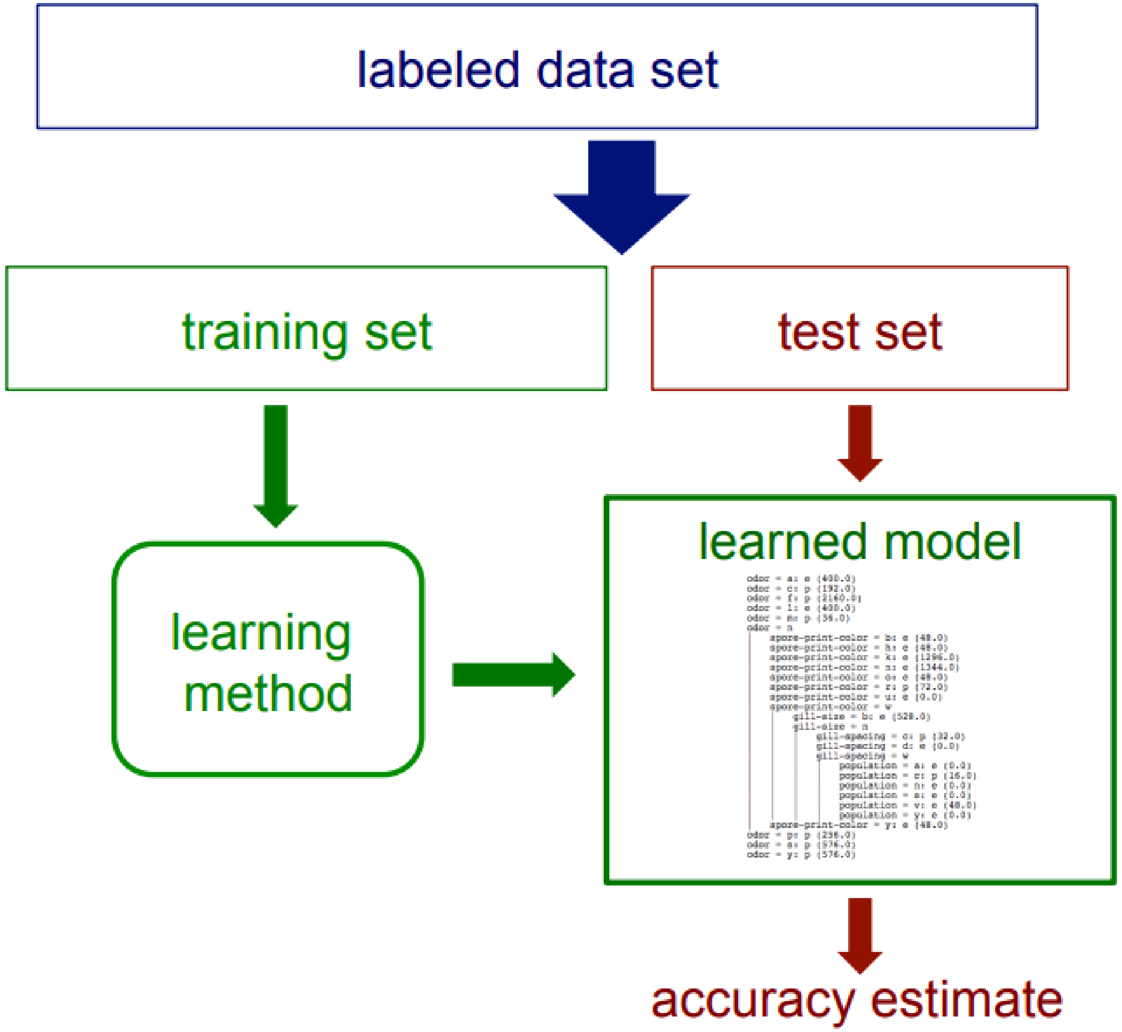
Hold-Off Evaluation

- How can we get an unbiased estimate of the accuracy of a learned model?
- Split data into two sets: training and test sets.
- Build model using the training set
- Evaluating model on the test set gives an estimation of generalization error



Test Set

- How can we get an unbiased estimate of the accuracy of a learned model?
- If the test-set labels influence the learned model in any way, accuracy estimates will be biased



Learning Curves

- How does the accuracy of a learning method change as a function of the training-set size?
- Plot accuracy against the training-set size.

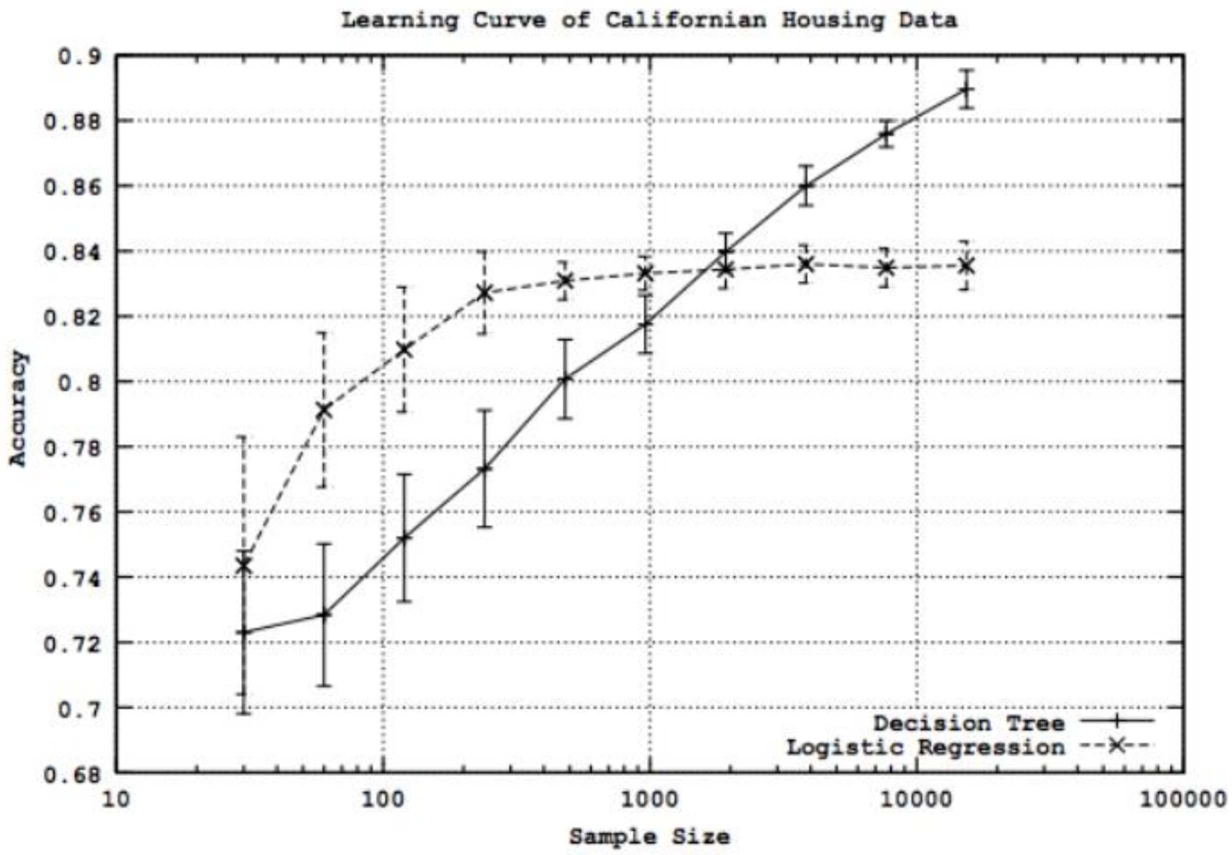


Figure from Perlich et al. Journal of Machine Learning Research, 2003

Learning Curves

- Given training/test set partition
- For each sample size s, randomly select *s* instances from training set
- Learn model
- Evaluate model on test set to determine accuracy *a*
- plot (*s*, *a*) or repeat n times and plot (*s*, avg. accuracy and error bars)

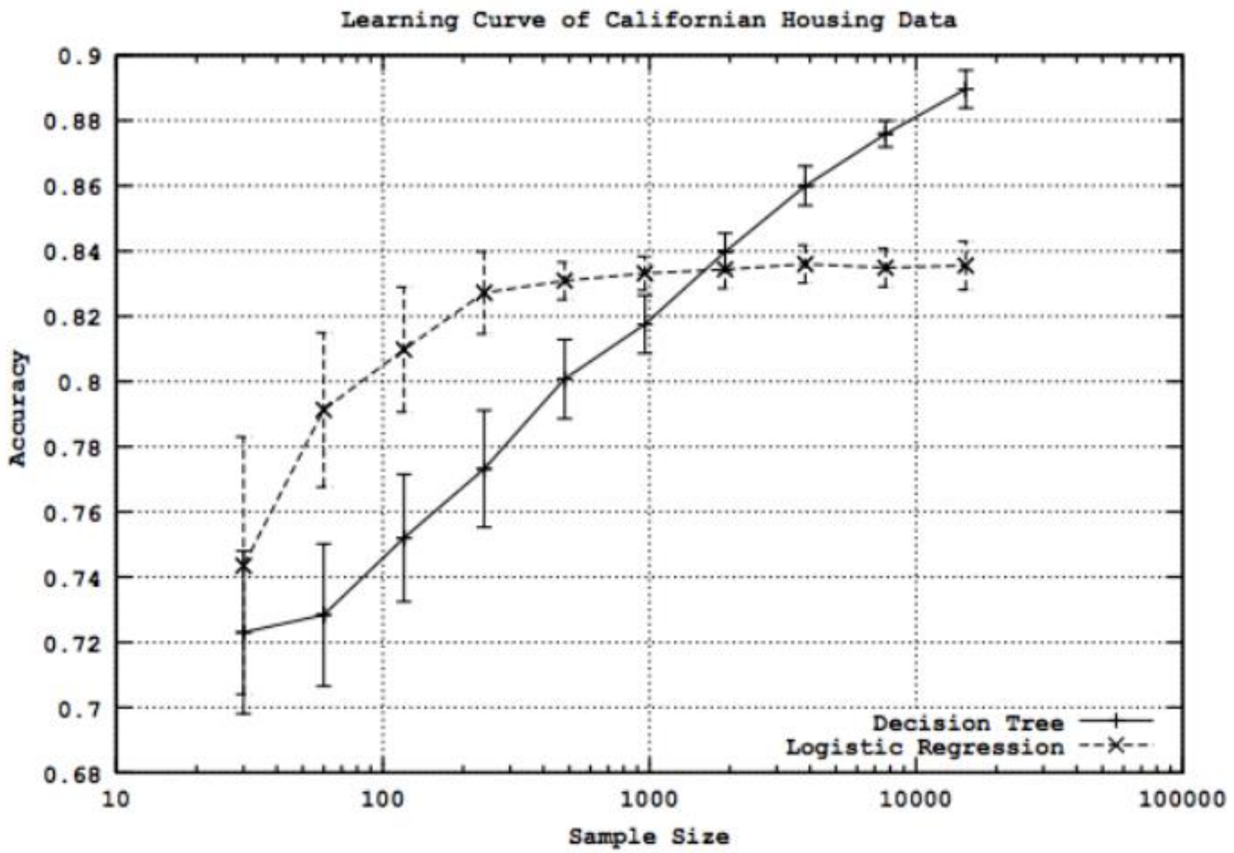
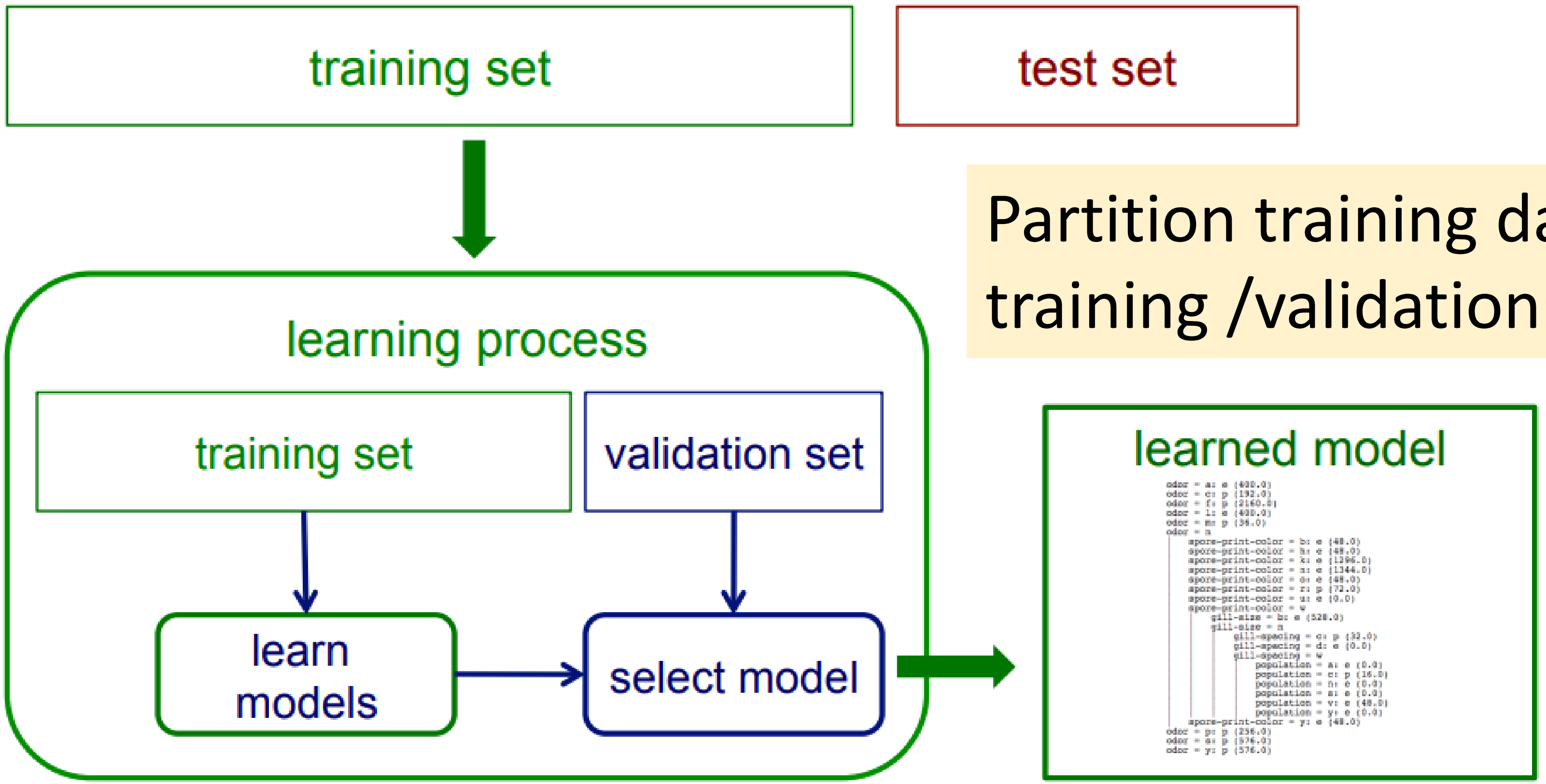


Figure from Perlich et al. Journal of Machine Learning Research, 2003

Validation Set

- Suppose we want unbiased estimates of accuracy during the learning process (e.g., to choose the best level of decision-tree pruning)?

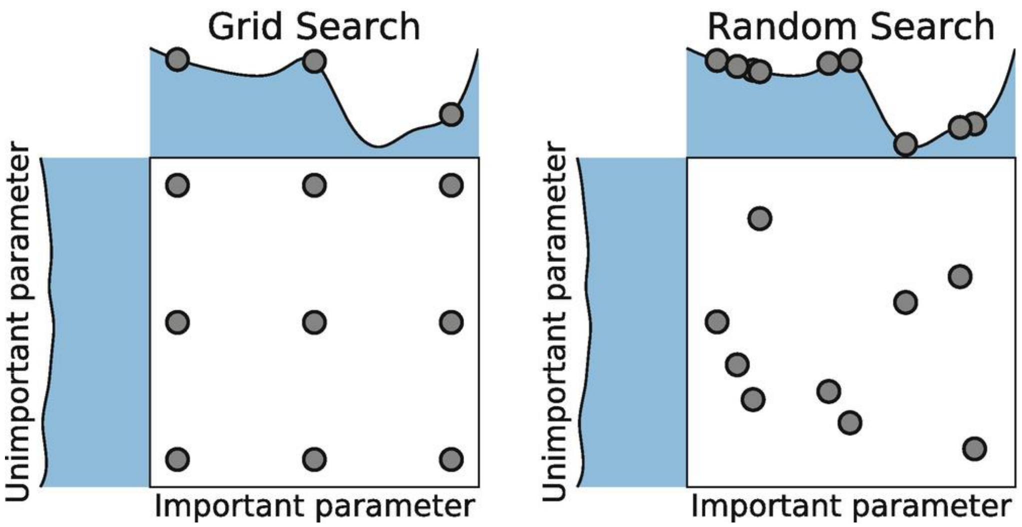


Partition training data into separate training /validation sets

Fine-Tune Your Model

- Find the best hyperparameter for the model

1. Grid Search = Search through all combinations of the parameters. (Scikit-Learn’s GridSearch)
2. Randomized Search = evaluates a preset number of random combinations. (Scikit-Learn’s RandomizedSearchCV)



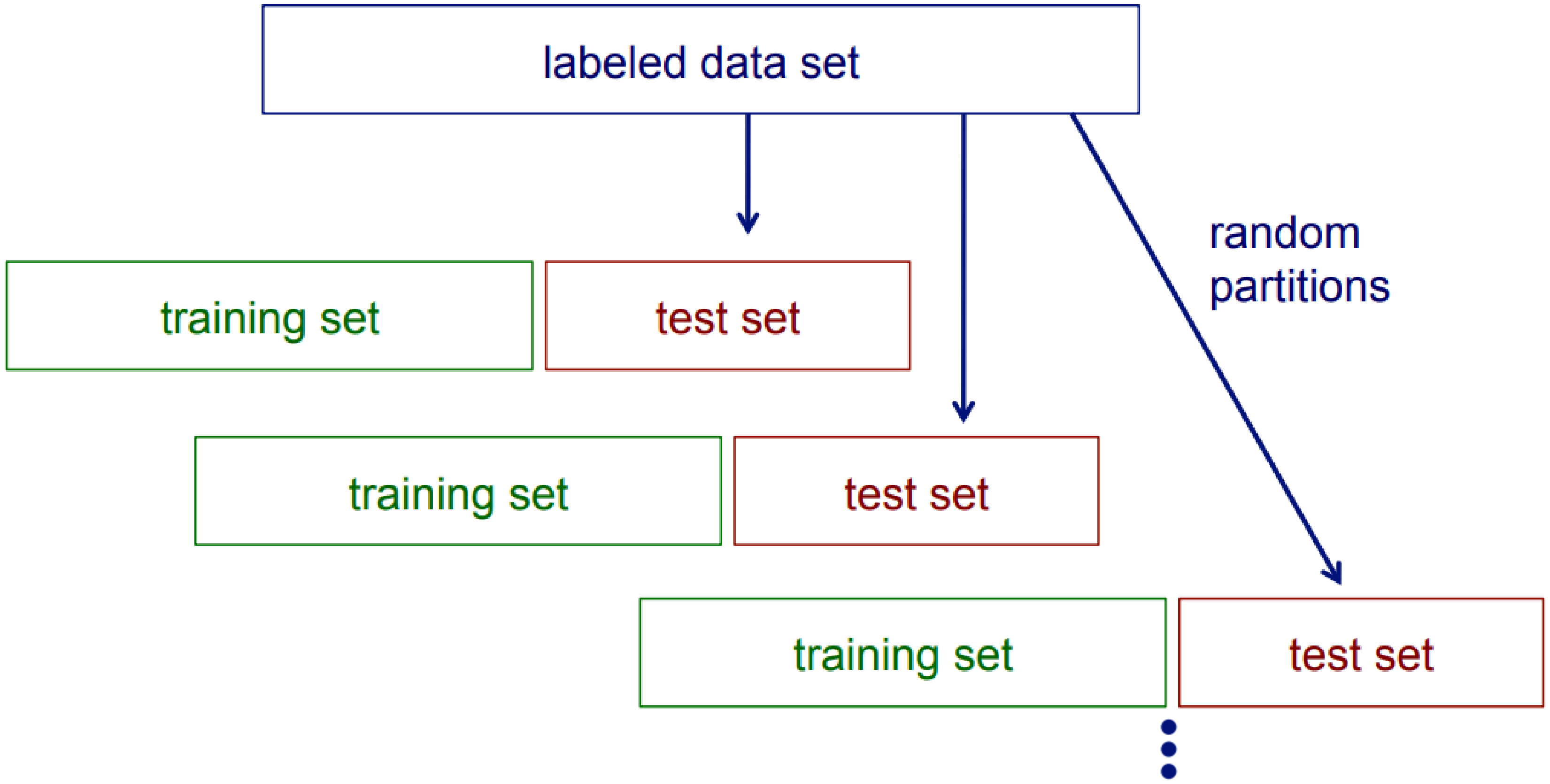
https://medium.com/@cjl2fv/an-intro-to-hyper-parameter-optimization-using-grid-search-and-random-search-d73b9834ca0a

Limitations of Using a Single Training/Test Partition

- We may not have enough data to make sufficiently large training and test sets
  - A larger test set gives us more reliable estimate of accuracy (i.e. a lower variance estimate)
  - But… a larger training set will be more representative of how much data we actually have for learning process
- A single training set doesn't tell us how sensitive accuracy is to a particular training sample

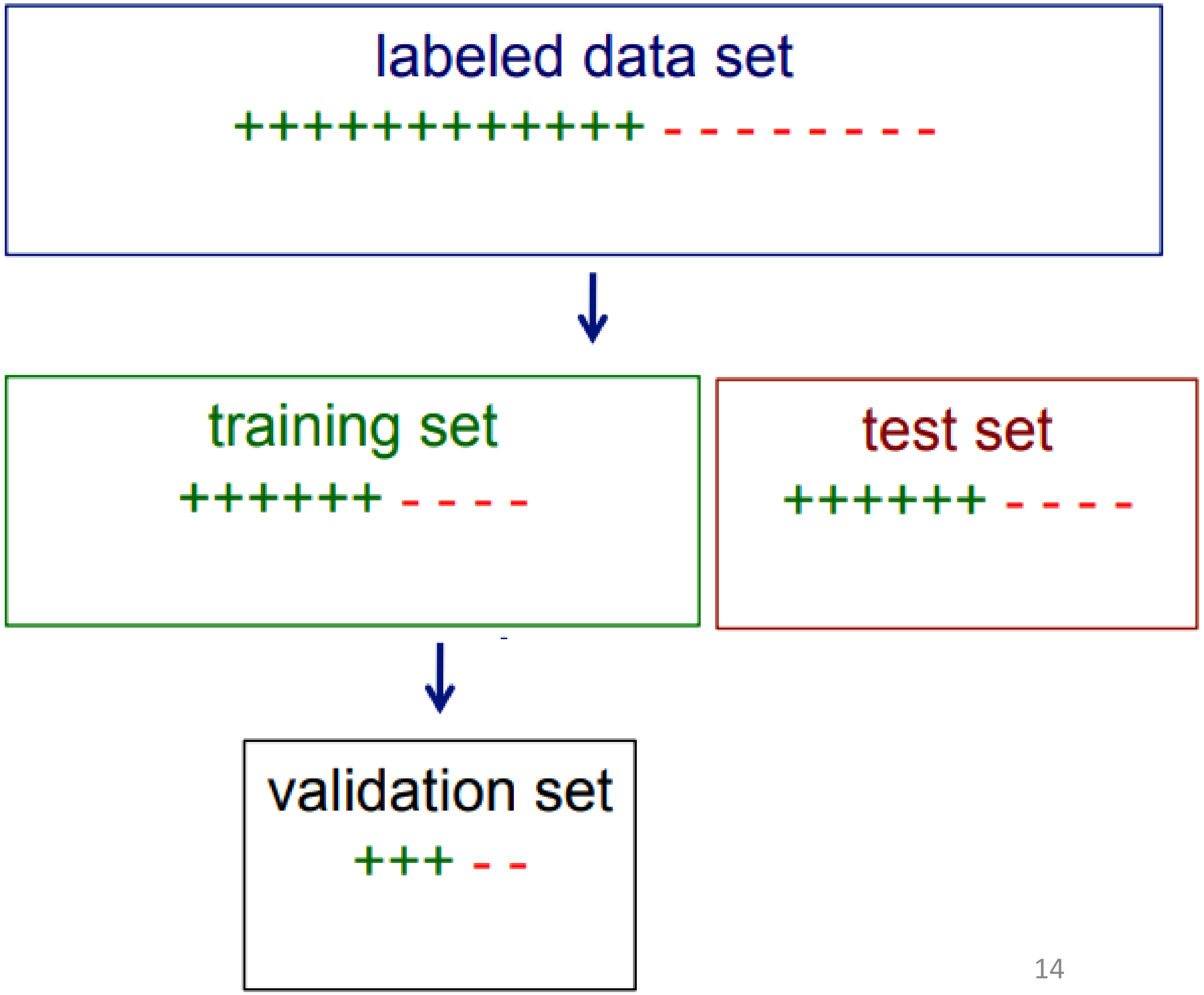
Random Resampling

- We can address the second issue by repeatedly randomly partitioning the available data into training and set sets.



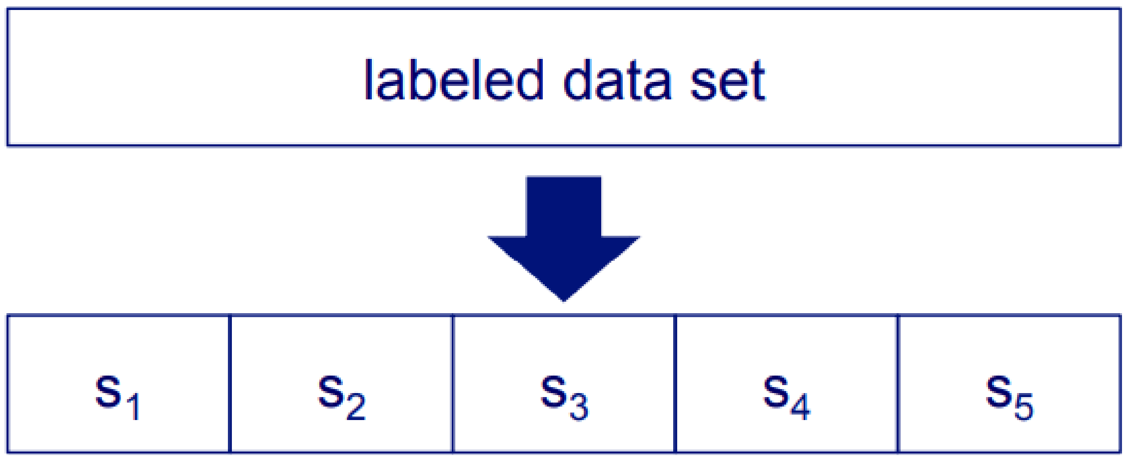
Stratified Sampling

- When randomly selecting training or validation sets, we may want to ensure that class proportions are maintained in each selected set.
- This can be done via stratified sampling: first stratify instances by class, then randomly select instances from each class proportionally



Cross Validation

- Partition data into n subsamples
- Iteratively leave one subsample out for the test set, train on the rest
- 10-fold CV is common,
- But smaller values of n are often used when learning takes time



iteration	train on	test on
1	s_2 s_3 s_4 s_5	s_1
2	s_1 s_3 s_4 s_5	s_2
3	s_1 s_2 s_4 s_5	s_3
4	s_1 s_2 s_3 s_5	s_4
5	s_1 s_2 s_3 s_4	s_5

Cross Validation Example

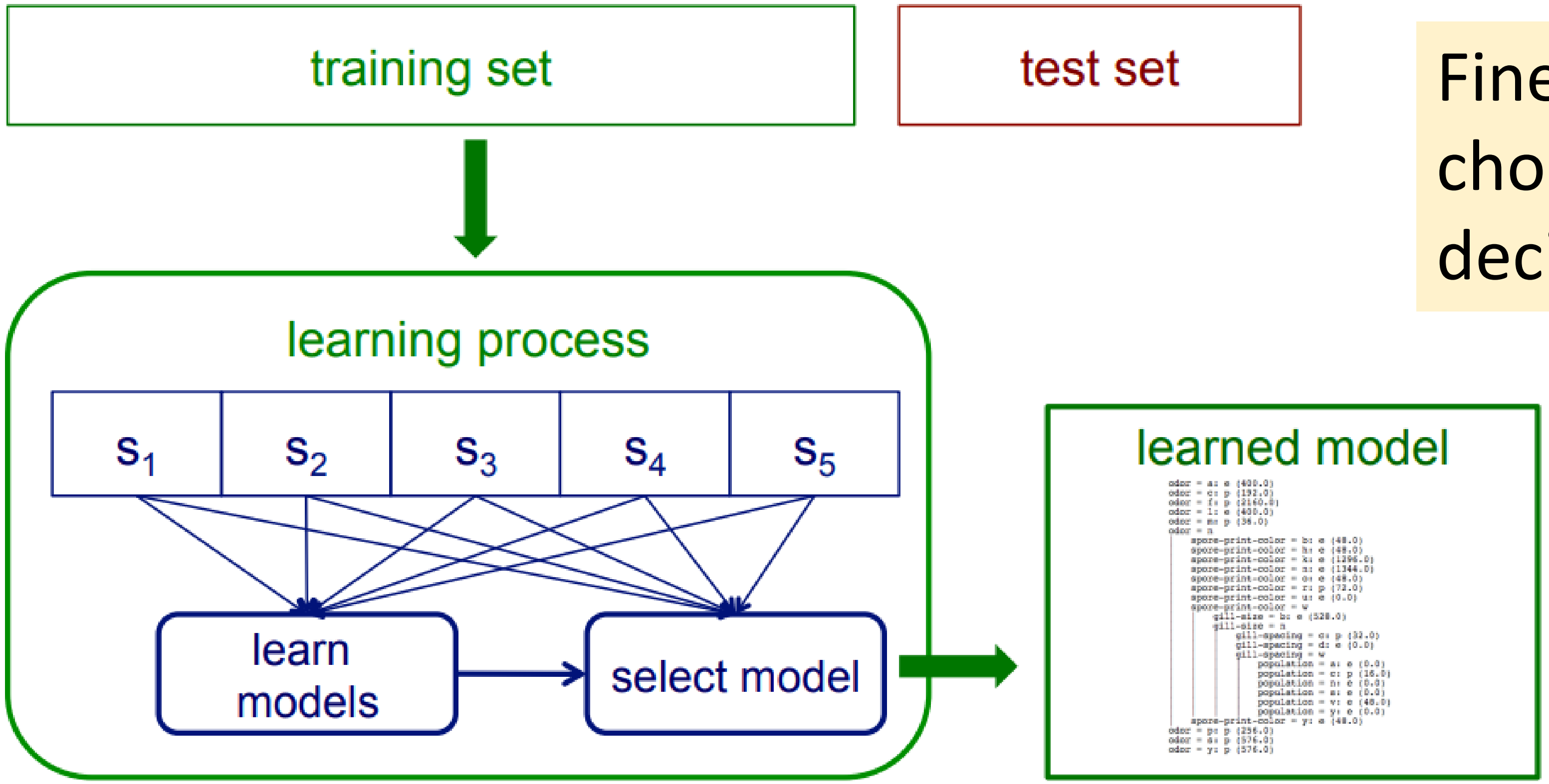
- Leave-one-out cross validation, n = # instances
- CV makes efficient use of the available data for testing
- Whenever we use multiple training sets, we are evaluating a learning method as opposed to an individual learned model.

iteration	train on	test on	correct
1	s_2 s_3 s_4 s_5	s_1	11 / 20
2	s_1 s_3 s_4 s_5	s_2	17 / 20
3	s_1 s_2 s_4 s_5	s_3	16 / 20
4	s_1 s_2 s_3 s_5	s_4	13 / 20
5	s_1 s_2 s_3 s_4	s_5	16 / 20

Accuracy = 73/100 = 73%

Internal Cross Validation

- Instead of a single validation set, we can use cross-validation within a training set to select a model



Fine tuning model, e.g., to choose the best level of decision-tree pruning

Performance Measures

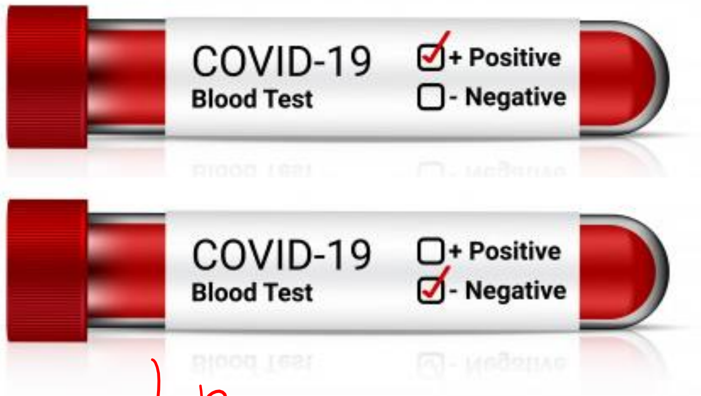
- Need a metric to measure/evaluate/compare classifiers and regressors.
- Evaluating classifier is trickier than evaluating a regressor.
  → real work is continuous, discrete attributes are what we defined.
- Metrics
  - **Classification:** Accuracy, F1-score, precision/recall, ROC curve
  - **Regression:** MAE, RMSE, R^2

Terminology

		actual class	
		positive	negative
predicted class	positive	true positives (TP)	false positives (FP)
	negative	false negatives (FN)	true negatives (TN)

pos

neg



	Pred	GT
TP	+	+
FP	+	−
FN	−	+
TN	−	−

Accuracy

- The most simple metric for classification
- correctly predicted samples / total testing samples

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

- Example:

Accuracy = 6/8 = 75%

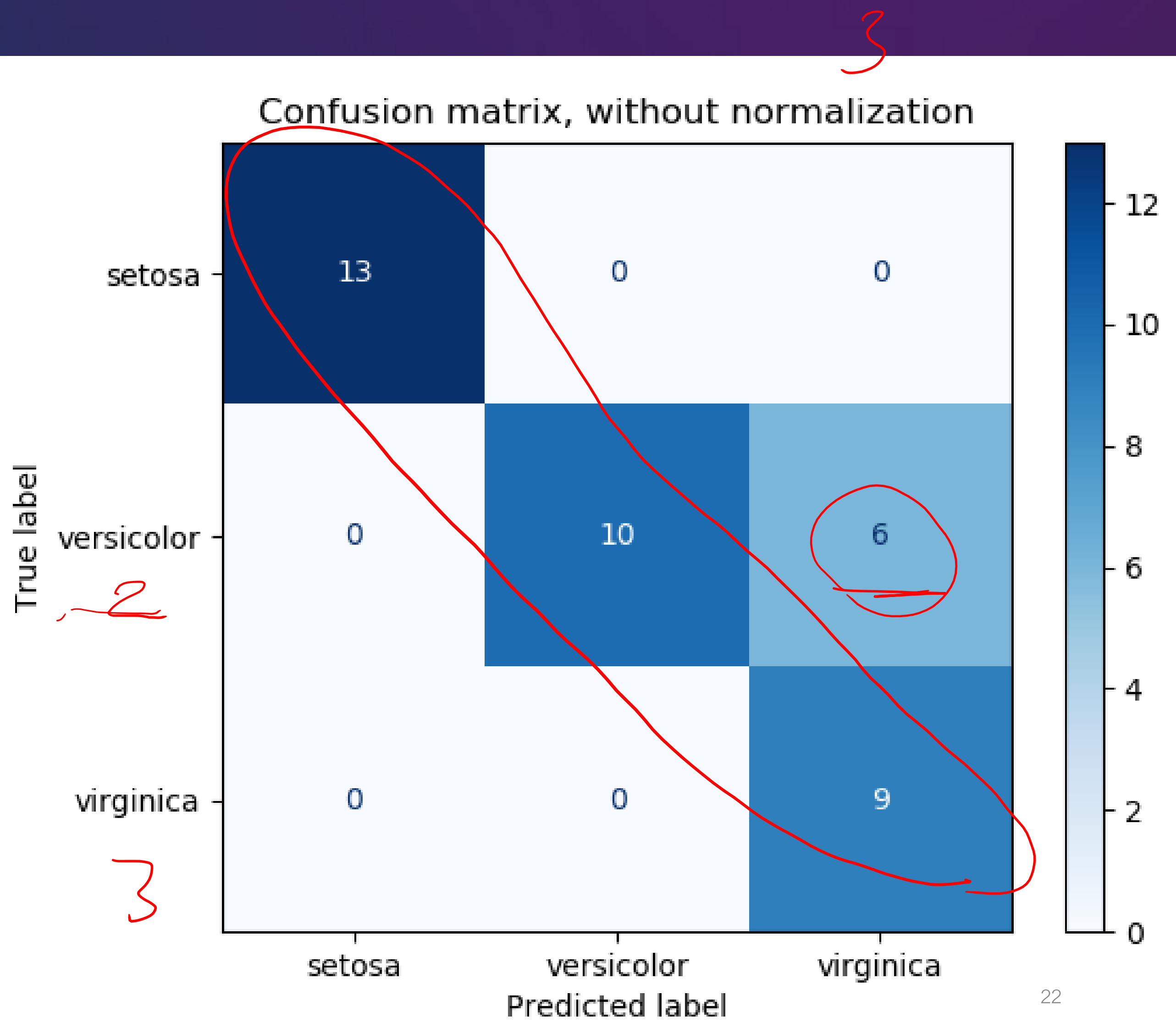
	Ground Truth	Prediction
1	Apple	Apple
2	Papaya	Papaya
3	Papaya	Papaya
4	Orange	Orange
5	Apple	Apple
6	Papaya	Orange
7	Apple	Apple
8	Banana	Papaya

Accuracy Issues

- Accuracy may not be useful measure in cases where:
  - ❑ There is a large class skew
    - ➢ Is 98% accuracy good if 97% of the instances are negative?
  - ❑ There are differential misclassification costs – say, getting a positive wrong costs more than getting a negative wrong
    - ➢ Consider a medical domain in which a false positive results in an extraneous test but a false negative results in a failure to treat a disease

Confusion Matrix

- Or error matrix
- A table whose entry (A,B) indicates the number of times instances of class A are predicted as class B.



Precision and Recall

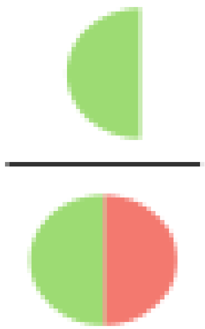
n = 100

Predict class 1 60

Ground Truth Class 1 40

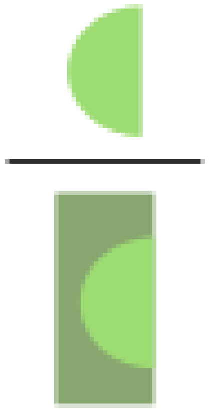
How many selected items are relevant?

$$\frac{TP}{TP+FP} = \text{Precision} =$$



How many relevant items are selected?

$$\text{Recall} =$$



$$= \frac{TP}{TP+FN}$$

https://en.wikipedia.org/wiki/Precision_and_recall

Example

Neg Class 1 99

Pos 2 1 P = 0 R = 0

Automated Diagnosis

- Automated system that diagnoses patients either he/she is unwell/normal TP
- Here, Positives = Unwell

Precision = $\frac{TP}{TP + FP} = \frac{1}{1+3} = 0.25$

Recall $= \frac{TP}{TP + FN} = \frac{1}{1+2} = 0.33$

	Ground Truth	Prediction
1	Unwell	Unwell
2	Unwell	Normal
3	Unwell	Normal
4	Normal	Normal
5	Normal	Unwell
6	Normal	Unwell
7	Normal	Unwell
8	Normal	Normal

F1-Score

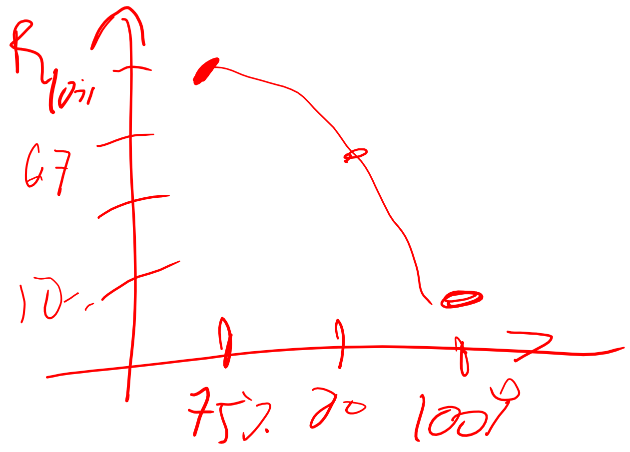
- F1 score = harmonic mean of precision and recall.

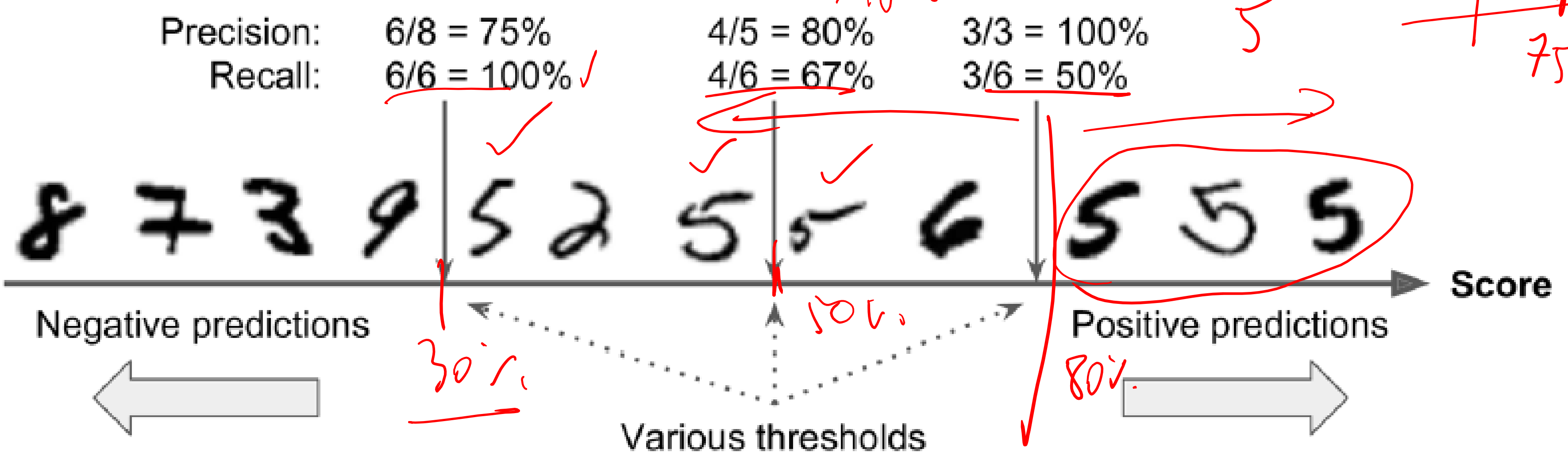
$$F_1 = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

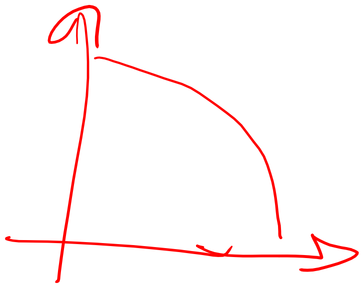
- The harmonic mean gives much more weight to low values.
  → Give high F1 score if both recall and precision are high.

Precision/Recall Tradeoff

- Assume a classifier is fixed and it assigns a score to each sample
- > threshold = positive, < threshold = negative
- High threshold = high precision, low recall
- Unfortunately, you can’t have it both ways.







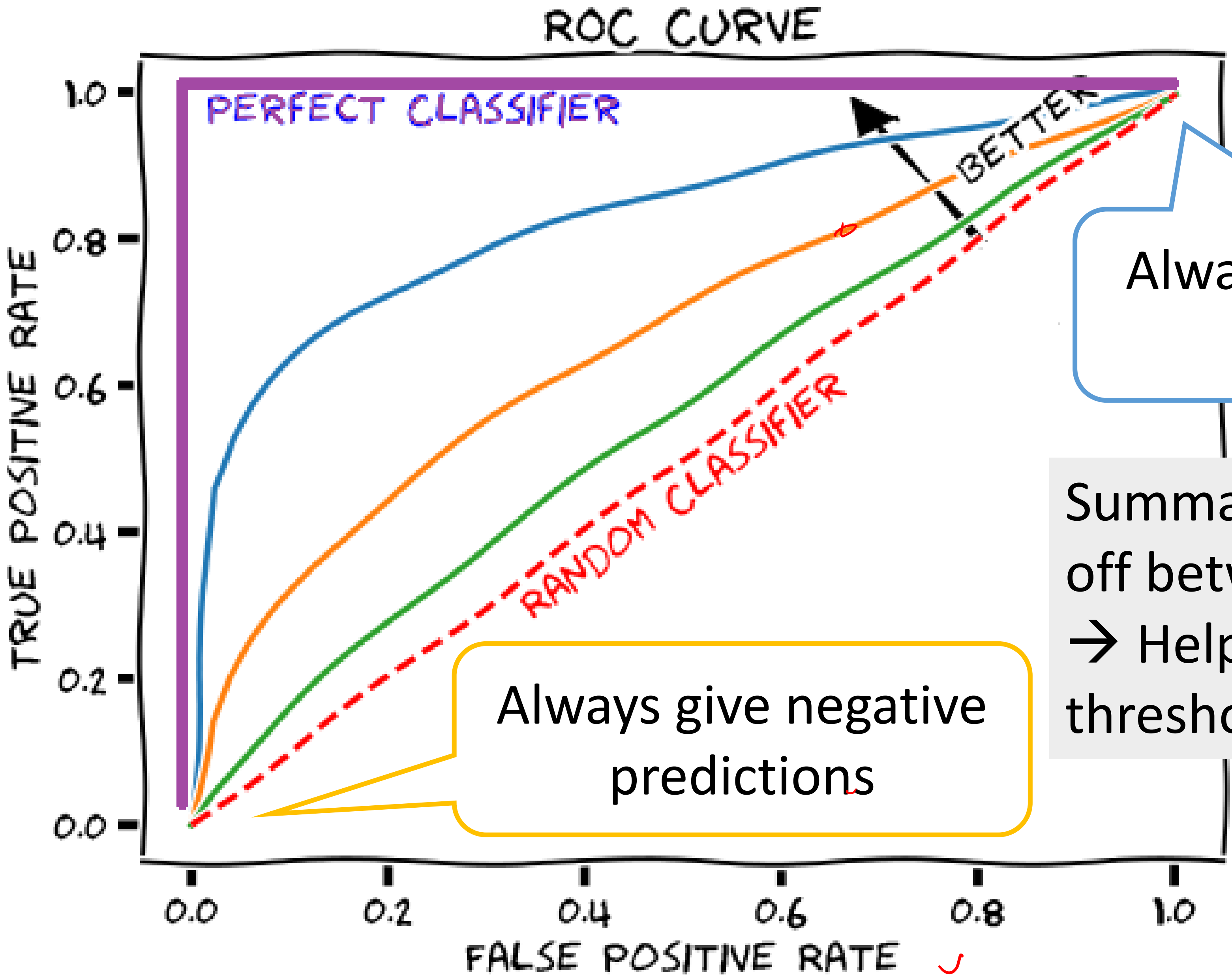
ROC Curve

- Receiver Operating Characteristic (ROC) curve
- For binary classifiers
- Plots the true positive rate (another name for recall) against the false positive rate for various threshold values.

$$\underbrace{\mathrm{TPR}}_{\text{recall}} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \qquad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{N}} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

- TPR = probability of detection
- FPR = probability of false alarm
- http://arogozhnikov.github.io/RocCurve.html

Recall



Always give positive predictions

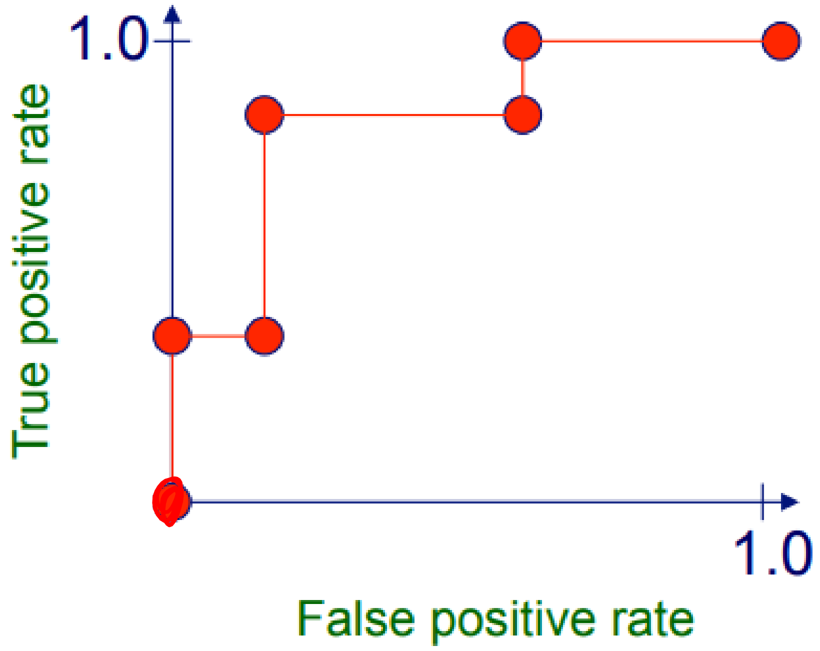
Always give negative predictions

Summarize the trade-off between TPR & FPR
→ Help determine the threshold

https://glassboxmedicine.com/2019/02/23/measuring-performance-auc-auroc/

ROC Curve Example

	instance	confidence positive		correct class
๑๐๐%				
	Ex 9	.99		+
	Ex 7	.98	TPR= 2/5, FPR= 0/5	+
๙๐%				
	Ex 1	.72	TPR= 2/5, FPR= 1/5	-
๗๐%				
	Ex 2	.70		+
	Ex 6	.65	TPR= 4/5, FPR= 1/5	+
๖๐%				
	Ex 10	.51		-
	Ex 3	.39	TPR= 4/5, FPR= 3/5	-
๓๐%				
	Ex 5	.24	TPR= 5/5, FPR= 3/5	+
๒๐%. ๑๐%				
	Ex 4	.11		-
	Ex 8	.01	TPR= 5/5, FPR= 5/5	-
๐%				



AUC

- To compare classifiers, we need a number.
  → Measure the area under the curve (AUC) of ROC.
- A perfect classifier will have ROC AUC = 1
- A purely random classifier will have ROC AUC = 0.5.
- To compute, one possible solution similar to Trapezoidal rule, divide regions at thresholds.

What about multiclass problem?

Demo: https://colab.research.google.com/drive/1u9AEoqBkhZDtyvI82OgG4jskN_9qkF2L?usp=sharing

Multilabel Classification

- Classification task labelling each sample with *x* labels from n_classes possible classes, where x can be 0 to n_classes inclusive.
- Thus, comparable to running n_classes binary classification tasks
- Multilabel classifiers may treat the multiple classes simultaneously, accounting for correlated behavior among them.
- E.g., In diagnosis system, one patient may have multiple diseases.

https://scikit-learn.org/stable/modules/multiclass.html

RMSE

- A typical performance measure for regression problems.
- Measure the standard deviation of the errors the system makes in its predictions

$$\mathrm{RMSE}(\mathbf{X}, h) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(h\left(\mathbf{x}^{(i)}\right) - y^{(i)}\right)^2}$$

- y = the expected output,
- X = data, m = no. of testing data and
- h = prediction function

MAE

- Or Average Absolute Deviation
- Another popular performance measure for regression problems

$$\mathrm{MAE}(\mathbf{X}, h) = \frac{1}{m} \sum_{i=1}^{m} \left| h\left(\mathbf{x}^{(i)}\right) - y^{(i)} \right|$$

- Since errors are squared in RMSE so it gives a relatively high weight to large errors
  → MAE is better when outliers are expected.

R^2

- Or coefficient of determination
- How well observed outcomes are replicated by the model → The proportion of total variation of outcomes explained by the model.

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

Total sum of squares

$$SS_{\text{tot}} = \sum_i (\underbrace{y_i}_{\text{Observed data}} - \underbrace{\bar{y}}_{\text{Mean of the observed data}})^2$$

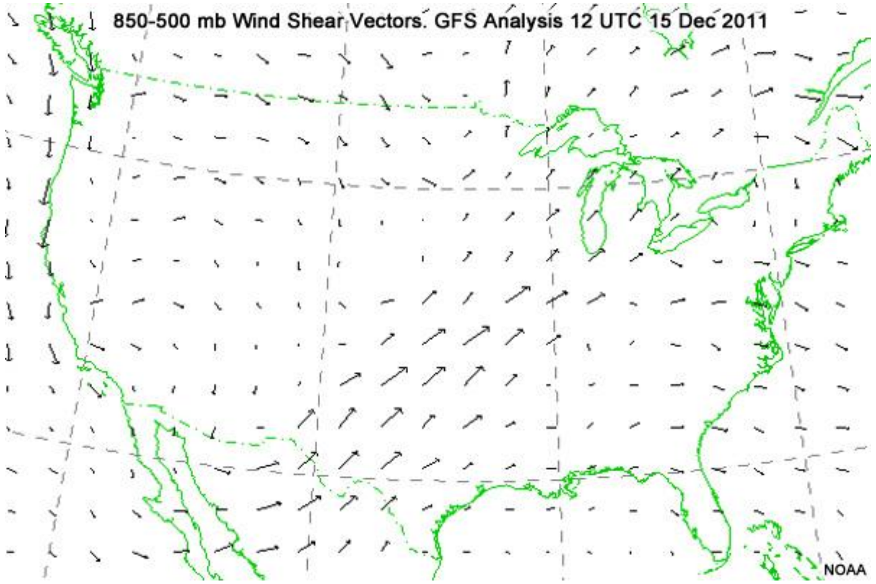
Residual sum of squares

$$SS_{\text{res}} = \sum_i (y_i - \underbrace{f_i}_{\text{Predicted value}})^2 = \sum_i e_i^2$$

https://en.wikipedia.org/wiki/Coefficient_of_determination

Multioutput Regression

- Predicts multiple numerical properties for each sample.
- Thus comparable to running n_output regression estimators
- Estimators that support multioutput regression may be faster.
- E.g., prediction of both wind speed and wind direction, in degrees, using data obtained at a certain location.
- Each sample would be data obtained at one location and both wind speed and direction would be output for each sample.



http://kejian1.cmatc.cn/vod/comet/dynamics/thermal_wind/navmenu.php_tab_1_page_11.2.1_type_text.htm

https://scikit-learn.org/stable/modules/multiclass.html

Comparing Learning Systems

- How can we determine if one learning system provides better performance than another
- Example:

	Accuracies on test sets				
System 1:	80%	50	75	…	99
System 2:	79	49	74	…	98
δ :	+1	+1	+1	…	+1

- Notice that System 1 is always better than System 2

Hypothesis Testing

- Hypothesis testing (or test of significance) is a statistical method used to determine if there is enough evidence in a sample data to draw conclusions about a population.
- Given a claim, identify the
  - The null hypothesis (H_0) = the value of a population parameter (e.g., mean, or SD) is equal to some claimed value.
  - The alternative hypothesis (H_a) the parameter differs from H_0
- Standard test for testing the difference between population: paired t-test and sign test

Comparing Systems Using a Paired t-Test

- consider δ's as observed values of a set of i.i.d. random variables
- H_0 : the 2 learning systems have the same accuracy
- H_a : one of the systems is more accurate than the other
- Use paired t-test to compute p-value or significance level that mean of δ's would arise from null hypothesis
- If p-value is sufficiently small (typically < 0.05) then reject the null hypothesis

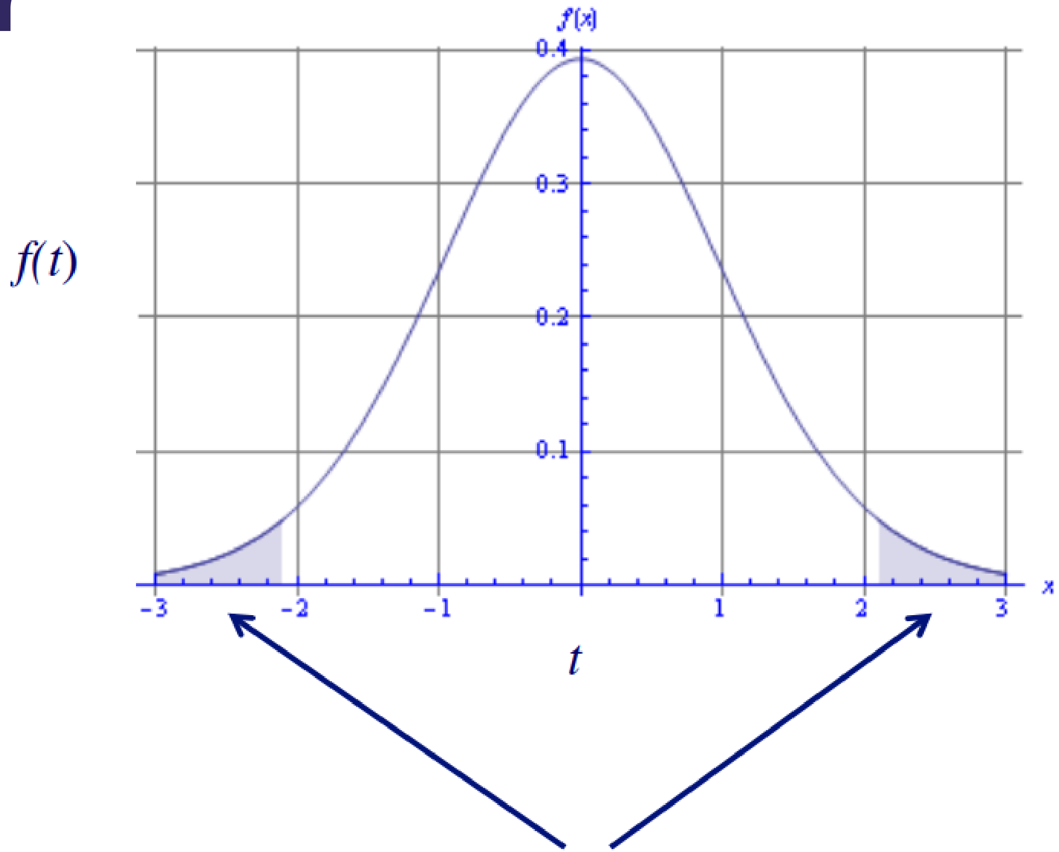
Comparing Systems Using a Paired t-Test

- Calculate the t statistic
- Determine the corresponding p-value, by looking up t in a table of values for the Student's t-distribution with n-1 degrees of freedom

$$t = \frac{\bar{\delta}}{\sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(\delta_i - \bar{\delta})^2}} \qquad \bar{\delta} = \frac{1}{n}\sum_{i=1}^{n}\delta_i$$

Comparing Systems Using a Paired t-Test

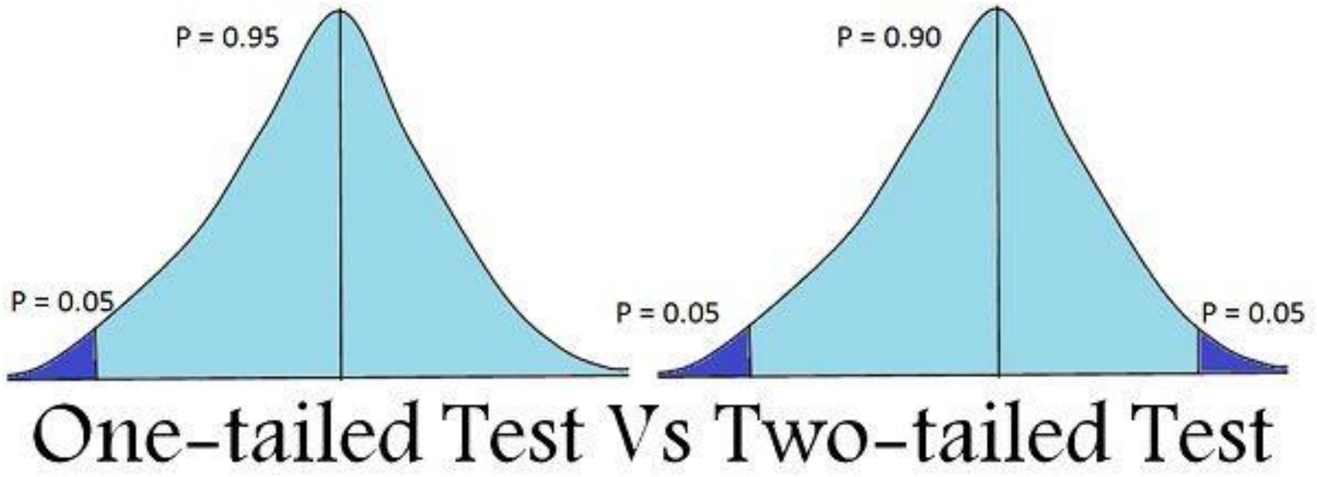
- The null distribution of our t statistic looks like this
- The p-value indicates how far out in a tail our t statistic is
- If the p-value is sufficiently small, we reject the null hypothesis, since it is unlikely we'd get such a value by chance



for a two-tailed test, the p-value represents the probability mass in these two regions

Why Do We Use a Two-Tailed Test?

- A two-tailed test asks the question: is the accuracy of the two systems different?
- A one-tailed test asks the question: is system A better than system B?
- A priori, we don't know which learning system will be more accurate (if there is a difference) – we want to allow that either one might be



Sign Test

- Test the null hypothesis that the median of a distribution is equal to some value.
- It is a non-parametric or “distribution-free” test = Don’t assume that data comes from a particular distribution, like the normal distribution.
- Count “wins” for Algorithm A and B over the N test examples on which they disagree
- Let M be the larger of these counts
- What is probability under b(N,0.5) that either A or B would win at least M times? → Find p-value from binomial table.

2.

Case Studies

Case Study : Kaggle

- Look at competition (maybe completed one) and notice:
  - Evaluation (Hold-out)
  - Metrics
  - How people get the solution (https://medium.com/kaggle-blog)
  - Interviews with Machine Learning Heroes (Kagglers + Practitioners + Researchers) (https://www.kaggle.com/discussions/general/76241)

OTTO – Multi-Objective Recommender System

- Build a multi-objective recommender system based on real-world e-commerce sessions
  → predict e-commerce clicks, cart additions, and orders.
- https://www.kaggle.com/competitions/otto-recommender-system
- Evaluation: Evaluated on the first 20 predictions for each action type, and the three recall values are weight-averaged:

$$score = 0.10 \cdot R_{clicks} + 0.30 \cdot R_{carts} + 0.60 \cdot R_{orders}$$

$$R_{type} = \frac{\sum_i^N |\{\text{predicted aids}\}_{i,type} \cap \{\text{ground truth aids}\}_{i,type}|}{\sum_i^N \min(20, |\{\text{ground truth aids}\}_{i,type}|)}$$

American Express - Default Prediction

- Predict if a customer will default in the future
- https://www.kaggle.com/competitions/amex-default-prediction
- The evaluation metric (M) = the mean of two measures of rank ordering: Normalized Gini Coefficient (G) and default rate captured at 4% (D):

$$M = 0.5 \cdot (G + D)$$

Kaggle - LLM Science Exam

- Answer difficult science-based questions written by a Large Language Model (LLM).
- https://www.kaggle.com/competitions/kaggle-llm-science-exam
- Submissions are evaluated according to the Mean Average Precision @ 3 (MAP@3):

$$MAP@3 = \frac{1}{U} \sum_{u=1}^{U} \sum_{k=1}^{min(n,3)} P(k) \times rel(k)$$

- This competition uses a hidden test.

Store Sales - Time Series Forecasting

- Use machine learning to predict grocery sales
- https://www.kaggle.com/competitions/store-sales-time-series-forecasting
- The evaluation metric for this competition is **Root Mean Squared Logarithmic Error**.

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\log(1+\hat{y}_i)-\log(1+y_i)\right)^2}$$

- How to split train-test data for time-series

NFL Big Data Bowl

- How many yards will an NFL player gain after receiving a handoff?
- https://www.kaggle.com/competitions/nfl-big-data-bowl-2020
- Submissions will be evaluated on the Continuous Ranked Probability Score (CRPS) → Predict the probability that the team gains <= that many yards on the play

$$C = \frac{1}{199N} \sum_{m=1}^{N} \sum_{n=-99}^{99} (P(y \leq n) - H(n - Y_m))^2$$

- **Solution motivation:** https://www.kaggle.com/c/nfl-big-data-bowl-2020/discussion/119400

Exercise:
Binary Classification with a Software Defects Dataset

- https://www.kaggle.com/competitions/playground-series-s3e23
- Predict defects in C programs given various attributes about the code.
- Feature description: https://www.kaggle.com/datasets/semustafacevik/software-defect-prediction

```
% 7. Attribute Information:
%
%      1. loc             : numeric % McCabe's line count of code
%      2. v(g)            : numeric % McCabe "cyclomatic complexity"
%      3. ev(g)           : numeric % McCabe "essential complexity"
%      4. iv(g)           : numeric % McCabe "design complexity"
%      5. n               : numeric % Halstead total operators + operands
%      6. v               : numeric % Halstead "volume"
%      7. l               : numeric % Halstead "program length"
%      8. d               : numeric % Halstead "difficulty"
%      9. i               : numeric % Halstead "intelligence"
%     10. e               : numeric % Halstead "effort"
%     11. b               : numeric % Halstead
%     12. t               : numeric % Halstead's time estimator
%     13. lOCode          : numeric % Halstead's line count
%     14. lOComment       : numeric % Halstead's count of lines of comments
%     15. lOBlank         : numeric % Halstead's count of blank lines
%     16. lOCodeAndComment: numeric
%     17. uniq_Op         : numeric % unique operators
%     18. uniq_Opnd       : numeric % unique operands
%     19. total_Op        : numeric % total operators
%     20. total_Opnd      : numeric % total operands
%     21: branchCount     : numeric % of the flow graph
%     22. defects         : {false,true} % module has/has not one or more
%                                        % reported defects
```