Evaluating & Comparing Machine Learning Models

Dr. Fayyaz ul Amir Afsar Minhas

PIEAS Biomedical Informatics Research Lab

Department of Computer and Information Sciences

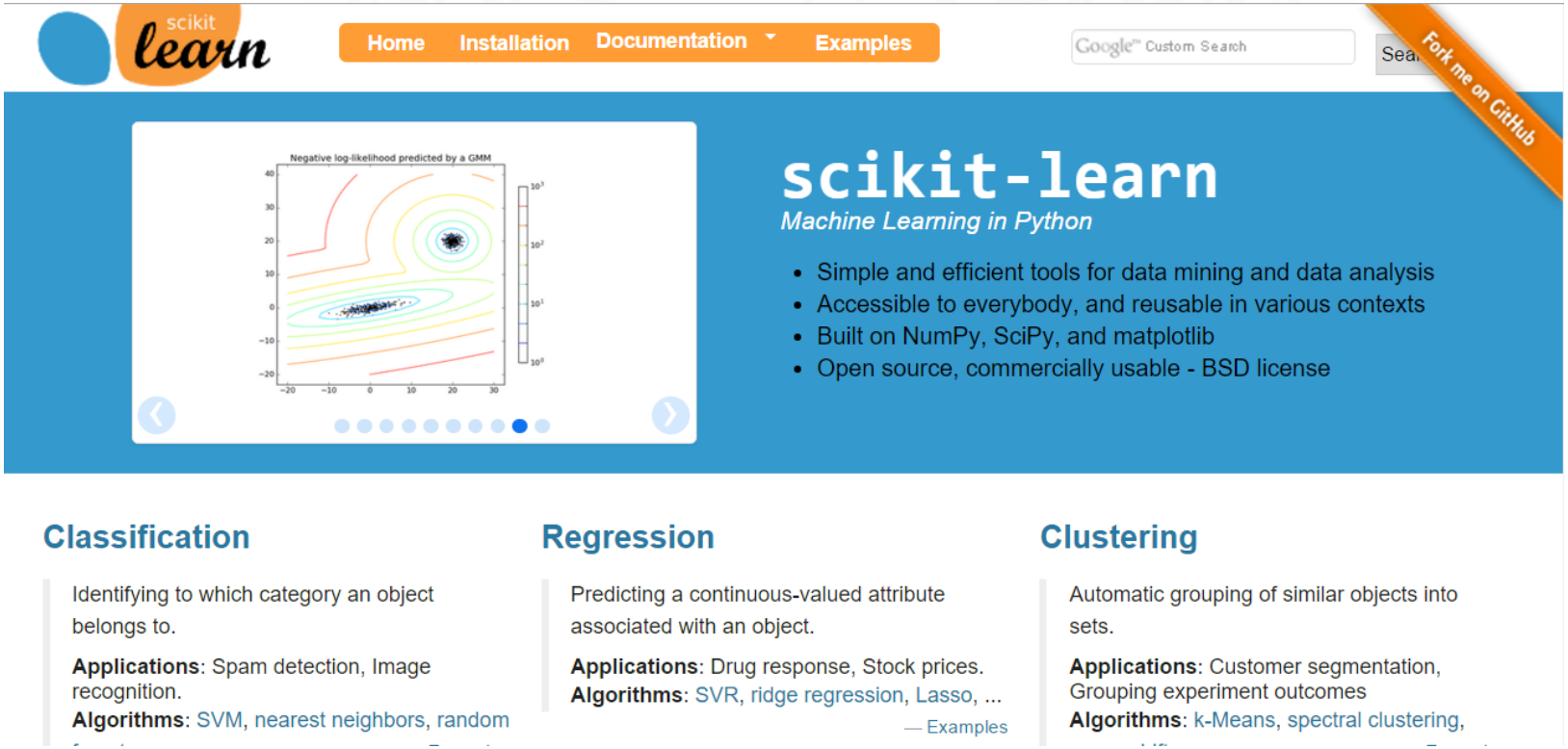
Pakistan Institute of Engineering & Applied Sciences

PO Nilore, Islamabad, Pakistan

http://faculty.pieas.edu.pk/fayyaz/

Introduction to scikit-learn

- http://scikit-learn.org/stable/tutorial/basic/tutorial.html
- Scikit-Learn SVM
  - http://scikit-learn.org/stable/modules/svm.html

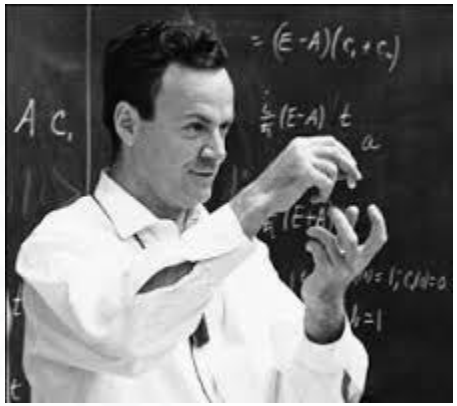


Objectives of this lecture

- How to compare machine learning models?
  - My classifier is better than yours
- How to select the optimal parameters of a machine learning model?
  - How should I choose “C” or “k”?
- Organization
  - **Philosophical Foundations**
  - **How to evaluate accuracy**
    - **Metrics: Accuracy, FPR, TRP, PPV, ROC, PR-Curves, F-measure**
    - **Cross-Validation and Resampling**
      - **K-fold, LOO, 5x2**
    - Bootstrapping
    - Interval Estimation
  - Experimental Design Strategies
  - Advanced Topics
    - Risk, Naïve Bayes, Structural Risk Minimization, etc.

Why Evaluate Models?

- Philosophical Foundations
- *"It doesn't matter how beautiful your theory is, it doesn't matter how smart you are. If it doesn't agree with experiment, it's wrong. "*
- Objective:
  - We want to find out which model "fits" our data best for use in the real world.

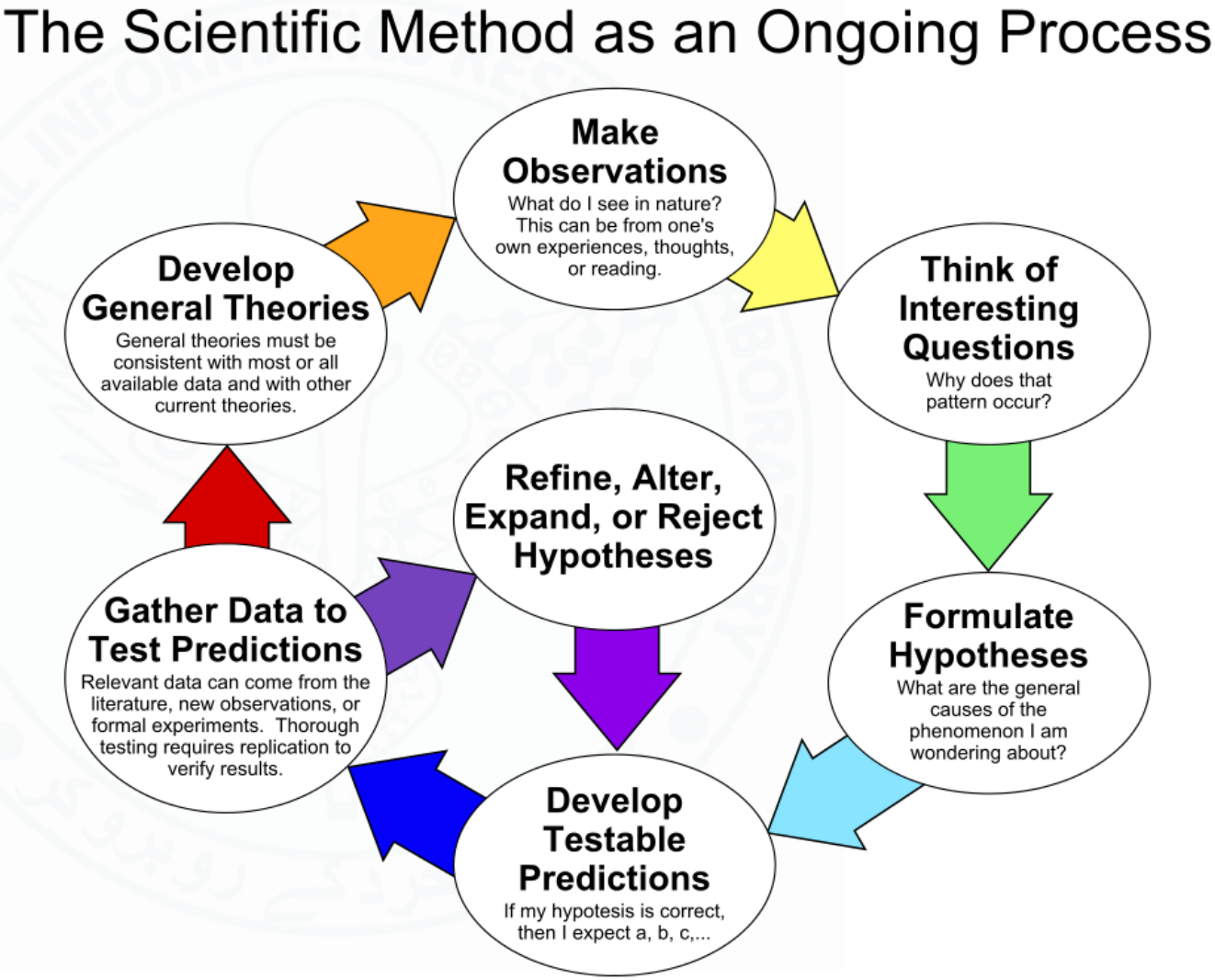


Richard Phillips Feynman 1918-1988

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

How do we compare classifiers?

- Comparison of machine learning models is and should be an exercise in the scientific method
- Requirements
  - Testability
  - Falsifiability
  - Experimental Observations
  - Reproducibility
  - Removal of Bias



https://en.wikipedia.org/wiki/Hypothetico-deductive_model

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

Scientific Method Summary

Feynman on Scientific Method

https://www.youtube.com/watch?v=ZswnT5uKn_E

- Guess
- Evaluate consequences
- Compare to Nature

- If it disagrees with nature, it's wrong!

A machine learning model

- Representation
- Evaluation
- Optimization
- Aims to find a sweet spot in the hypothesis space that generalizes / solves the problem

Domingos, Pedro. “A Few Useful Things to Know About Machine Learning.” Commun. ACM 55, no. 10 (October 2012): 78–87. doi:10.1145/2347736.2347755.

Evaluating & Comparing Models

- **Be clear about the objective of your evaluation**
  - I want to find the best parameters for my model
  - Is my model better on this data?
  - Is this classifier typically better than this other one?
  - Does this model work?
  - This model does not work because …
    - It fails to capture the structure of the data
    - It fails to work on discrete data
    - The data is not linearly separable
    - The amount of training data is small
    - The data is imbalanced
    - The training data is noisy
    - The test data does not follow the same distribution as the training data
    - The underlying assumptions of this classifier need revision
    - The **optimization** algorithm failed
    - The **representation** is improper
    - The **evaluation** strategy presented in the paper by Lay man et al. is wrong
  - This model gives better sensitivity
  - My training time is better than yours
  - The classifier is overfitting/ poor at generalization
  - This model is particular suited for high dimensional data
  - These features work better than these other features
  - When to stop learning?
  - Etc.

Experiment Design Objectives

- Accuracy Evaluation: How good will it be in practice?
- Sensitivity Analysis: Is the classifier sensitive to
  - the choice of the parameters so much so that it will be useless in practice
  - choice of the data
  - Randomness
  - Round-off error
  - Other controllable and non-controllable factors

Evaluation Metrics

- Training Set: For training the model
- Test Set: For evaluation
- Under no circumstances are testing labels to be used in training or the training data in evaluation of the generalization performance
- All evaluation metrics have underlying assumptions and limitations which may or may not suffice for the test that you are trying to perform
- Two-Class Classification
  - Accuracy?
    - Assumption
      - The data set is imbalanced
      - Misclassification of any class is equally bad?
      - The threshold used for classification will be used in practice

Classification Performance

- A classifier (or any machine learning model) can be viewed as a function $\mathrm{y} = f(x|\theta)$ which generates an output y given the input x and a parameter set θ using a decision function $f(x|\theta)$
- The output of a classifier is typically a real-valued output which is then thresholded to yield classification labels
  - $f(x|\theta) > 0 \Rightarrow y = +1$
  - $f(x|\theta) < 0 \Rightarrow y = -1$
- Here "0" acts as the threshold
- *What is the corresponding rule of the 1-NN classifier?*
  - *For the k-NN classifier?*
- Thus, the labels can change based on the threshold
- Thus, accuracy of a classifier is parametrized by this threshold

Thought Experiment

- Consider the data
  - All examples with $x < 0.5$ will be negative and the others will be positive
  - Assume that the data is balanced (equal number of positive and negative examples)
  - Consider a random classifier
    - This classifier will generate a random score of any example given as input
    - What will be its accuracy?
  - Consider a classifier which generates a score of +1 for all inputs
    - What will be its accuracy?

Name	Formula
error	$(fp + fn)/N$
accuracy	$(tp + tn)/N = 1 - \text{error}$
tp-rate	tp/p
fp-rate	fp/n
precision	tp/p'
recall	$tp/p = \text{tp-rate}$
sensitivity	$tp/p = \text{tp-rate}$
specificity	$tn/n = 1 - \text{fp-rate}$

		Patients with bowel cancer (as confirmed on endoscopy)		
		Condition positive	Condition negative	
Fecal occult blood screen test outcome	Test outcome positive	**True positive** (TP) = 20	**False positive** (FP) = 180	Positive predictive value = TP / (TP + FP) = 20 / (20 + 180) = **10%**
	Test outcome negative	**False negative** (FN) = 10	**True negative** (TN) = 1820	Negative predictive value = TN / (FN + TN) = 1820 / (10 + 1820) ≈ **99.5%**
		Sensitivity = TP / (TP + FN) = 20 / (20 + 10) ≈ **67%**	**Specificity** = TN / (FP + TN) = 1820 / (180 + 1820) = **91%**	

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

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

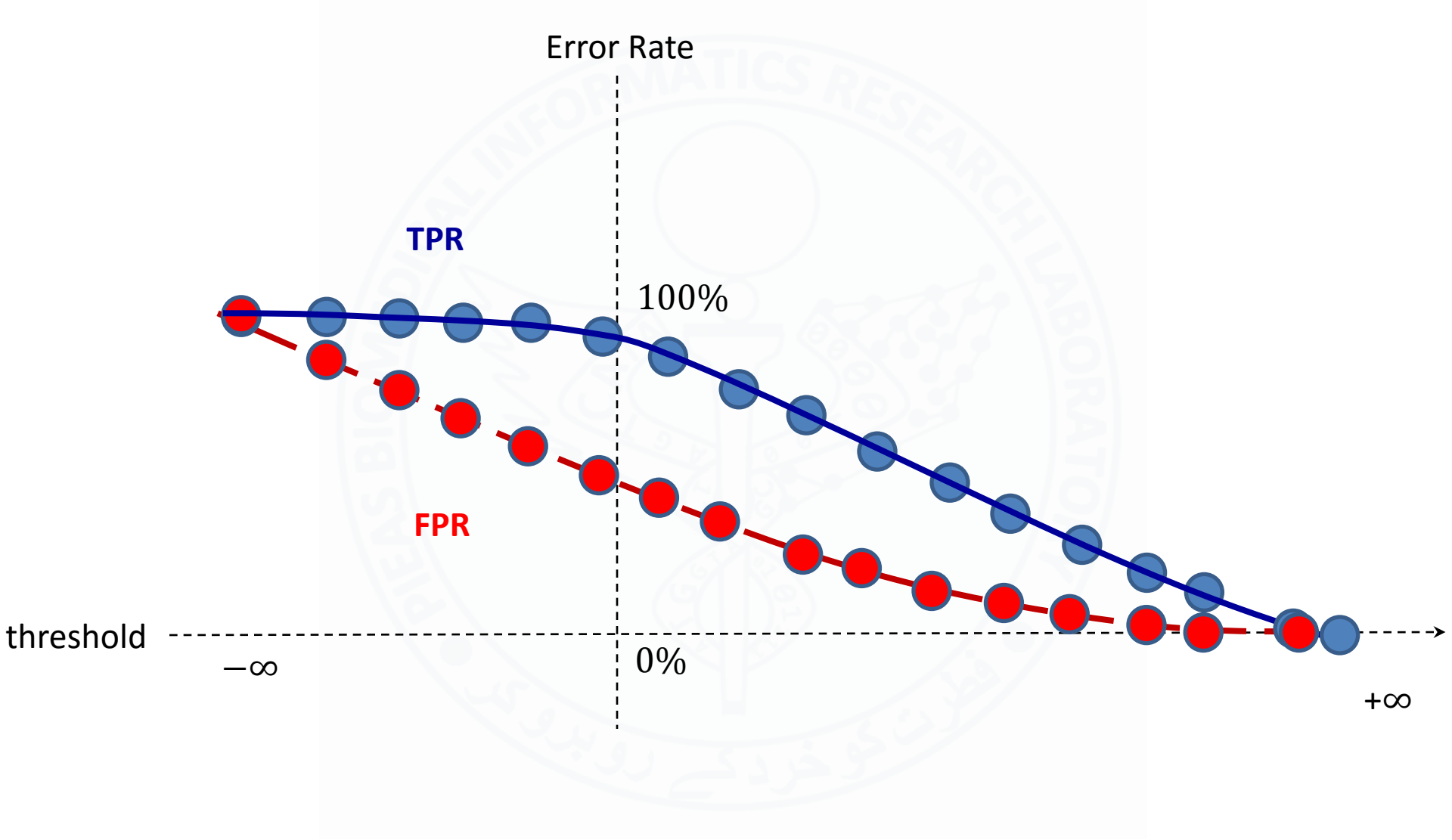
Confusion Matrix

		True condition			
	Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**	**False positive** (Type I error)	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$	False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Test outcome positive}}$
	Predicted condition negative	**False negative** (Type II error)	**True negative**	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Test outcome negative}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Test outcome negative}}$
	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	True positive rate (TPR), Sensitivity, Recall $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	True negative rate (TNR), Specificity (SPC) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$	

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

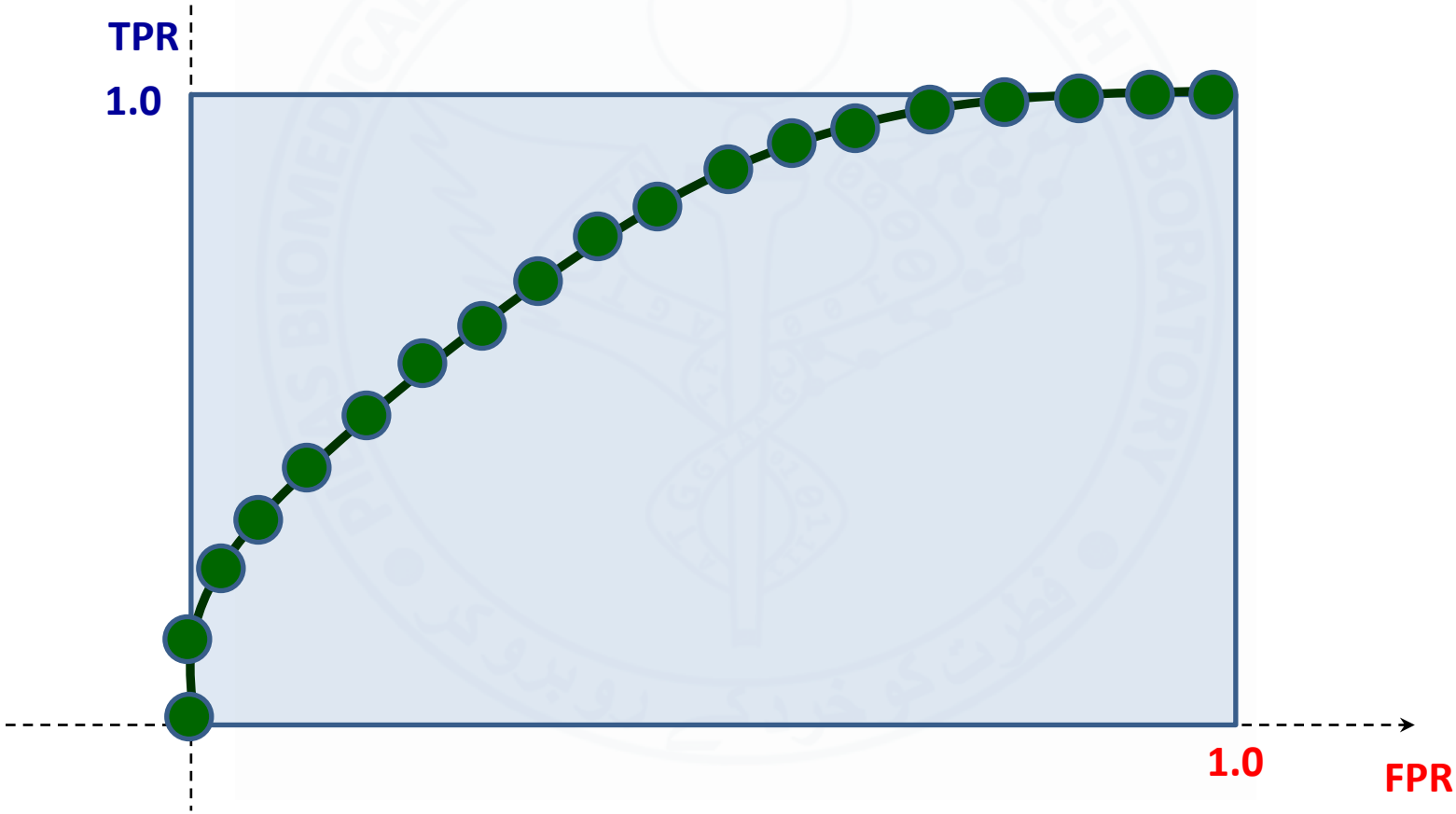
- What will be the behavior of TPR with increase in threshold of the classifier?
- How will FPR behave?
- How will Precision behave?
- Can TPR decrease with increase in threshold?

FPR vs. TPR Curve



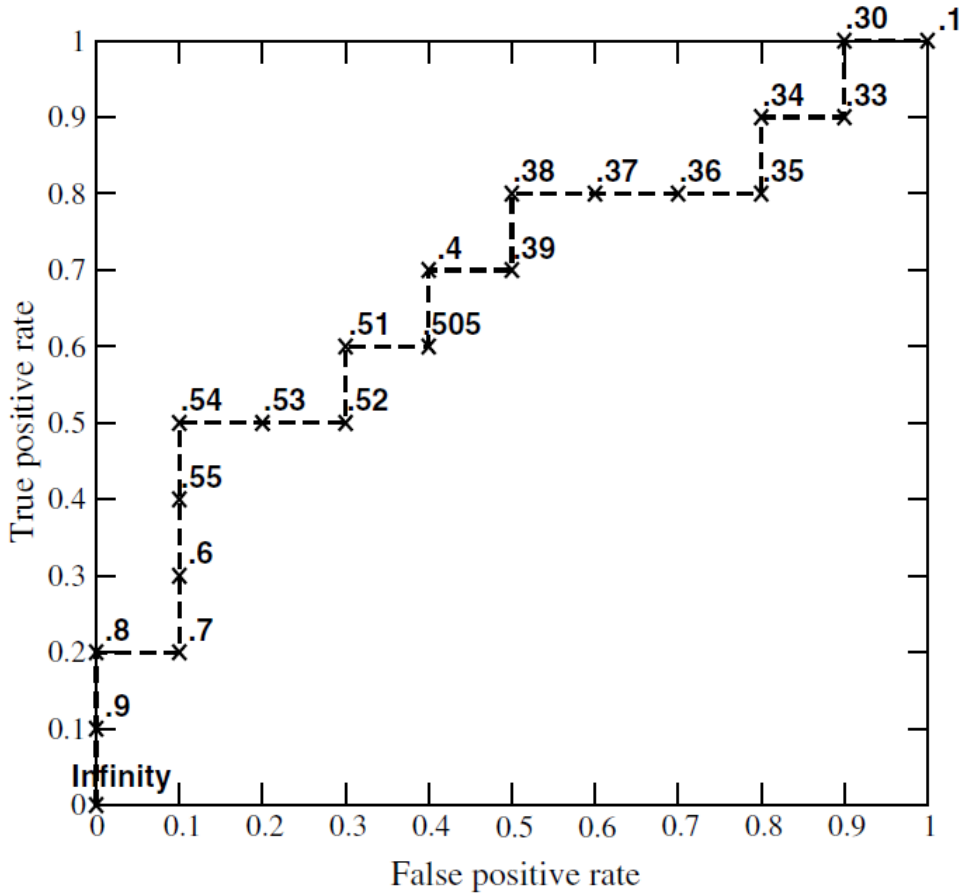
Receiver Operating Characteristics Curve

- A plot of TPR vs FPR



Making the ROC Curve

Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1

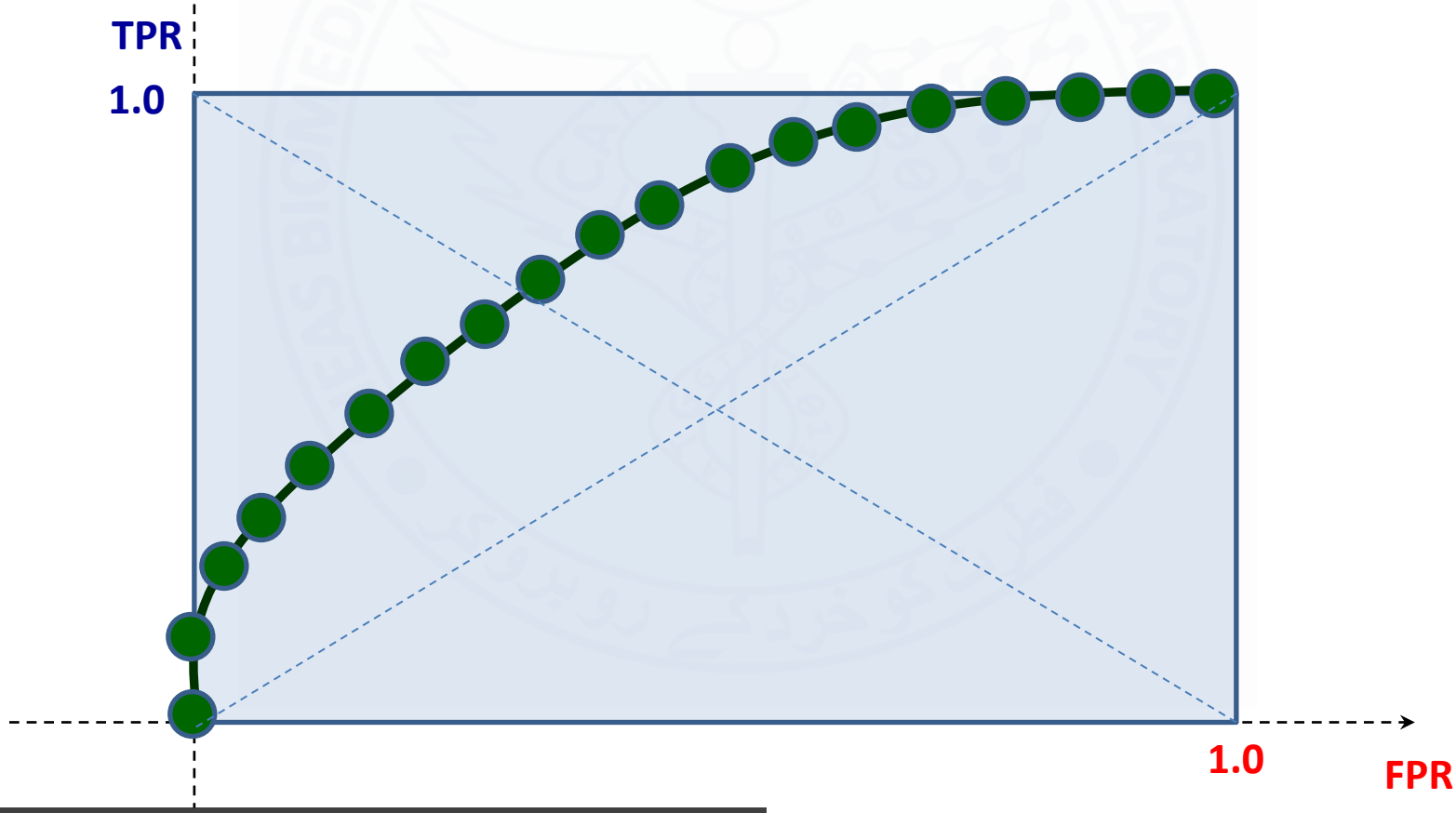


ROC

- What will be ROC curve for a perfect classifier?
- What will the ROC Curve of a random classifier look like?
- What will the ROC curve of a classifier that always predicts the positive class look like?
- What are the underlying assumptions of the ROC curve?
- What part of the ROC curve is the most important?

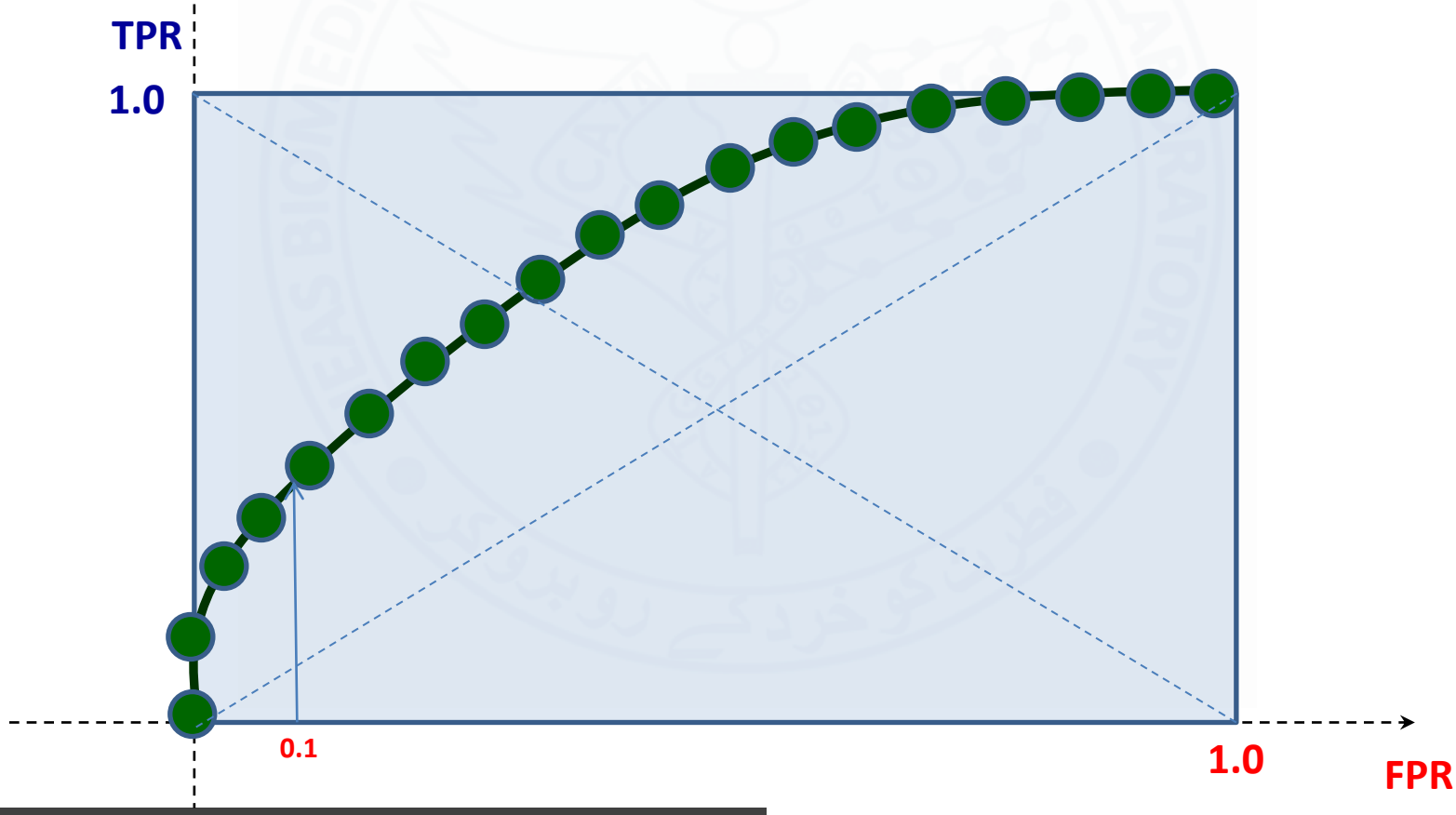
AUC-ROC

- The area under the ROC curve is a quality metric



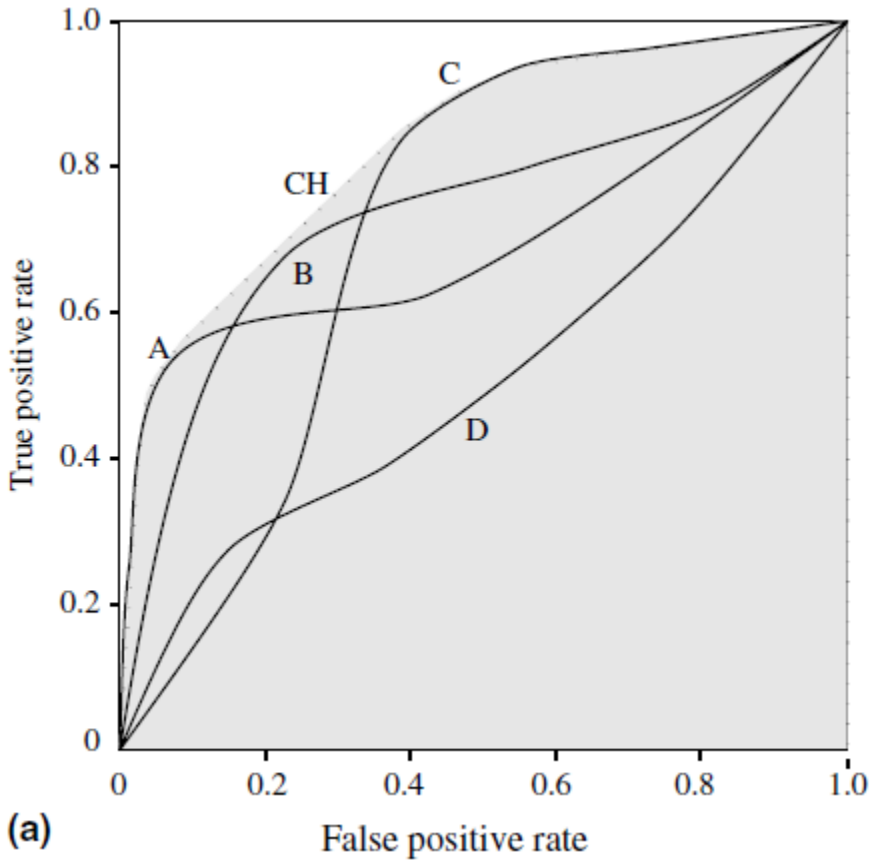
AUC ROC-N

- Area under the ROC curve up to the first N False Positives
  - N = 50
  - N = 10%

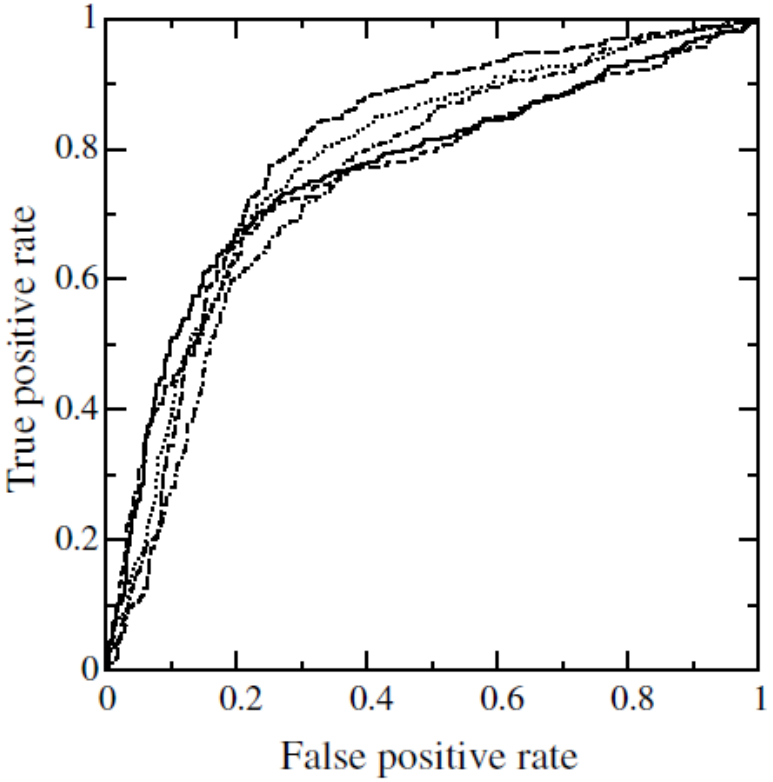


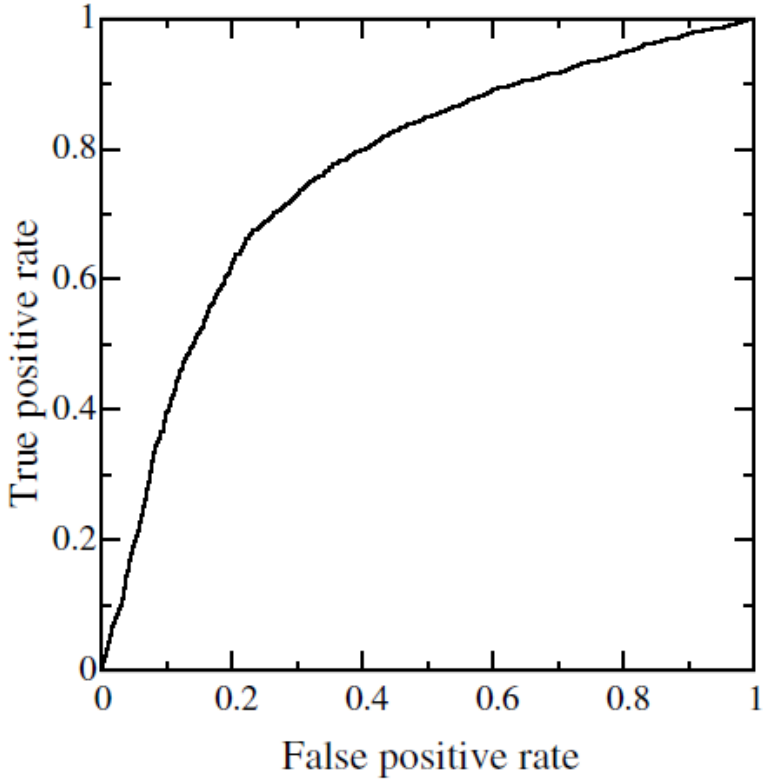
ROC Convex Hull

- Scores of two classifiers can be combined through a weighted combination to result in an optimal classifier
- This can be done using the ROC convex hull



Averaging ROC Curves





Multi-class ROC Curves

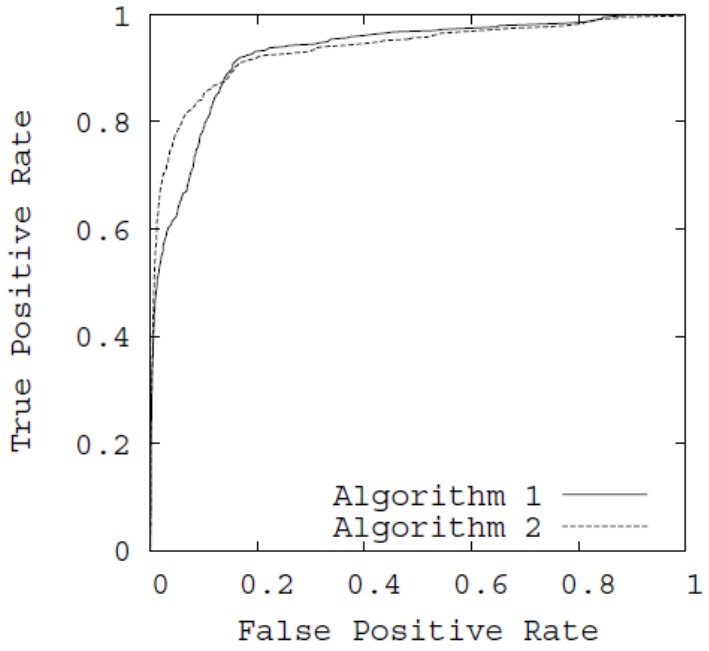
- Can also make multiple class ROC curves
  - One vs. Rest
- AUC-ROC can also be computed
  - Pairwise

Properties

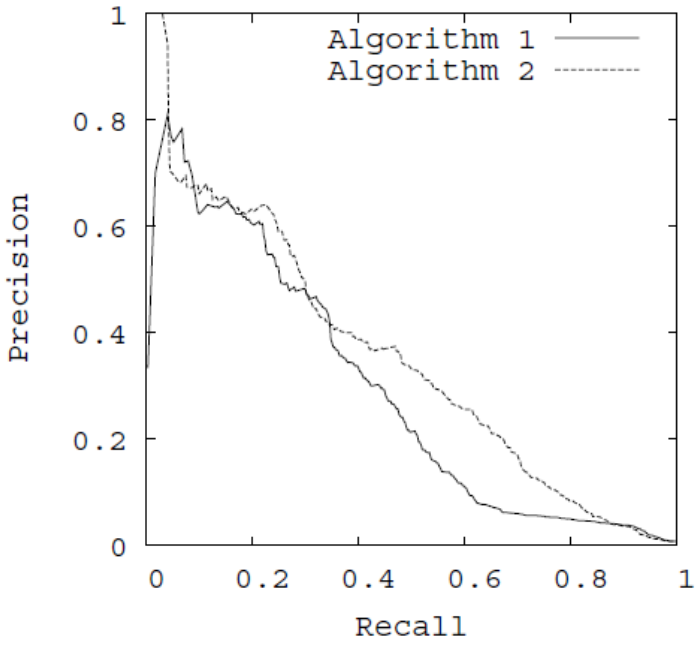
- Class Imbalance?
- When to Use?
- What to focus on?
  - FPR?
  - TPR?

Precision-Recall Curve

- Plot of Precision vs. Recall
- AUC-PR is a performance metric
- Useful in cases of class-imbalance or in which precision is a requirement



(a) Comparison in ROC space



(b) Comparison in PR space

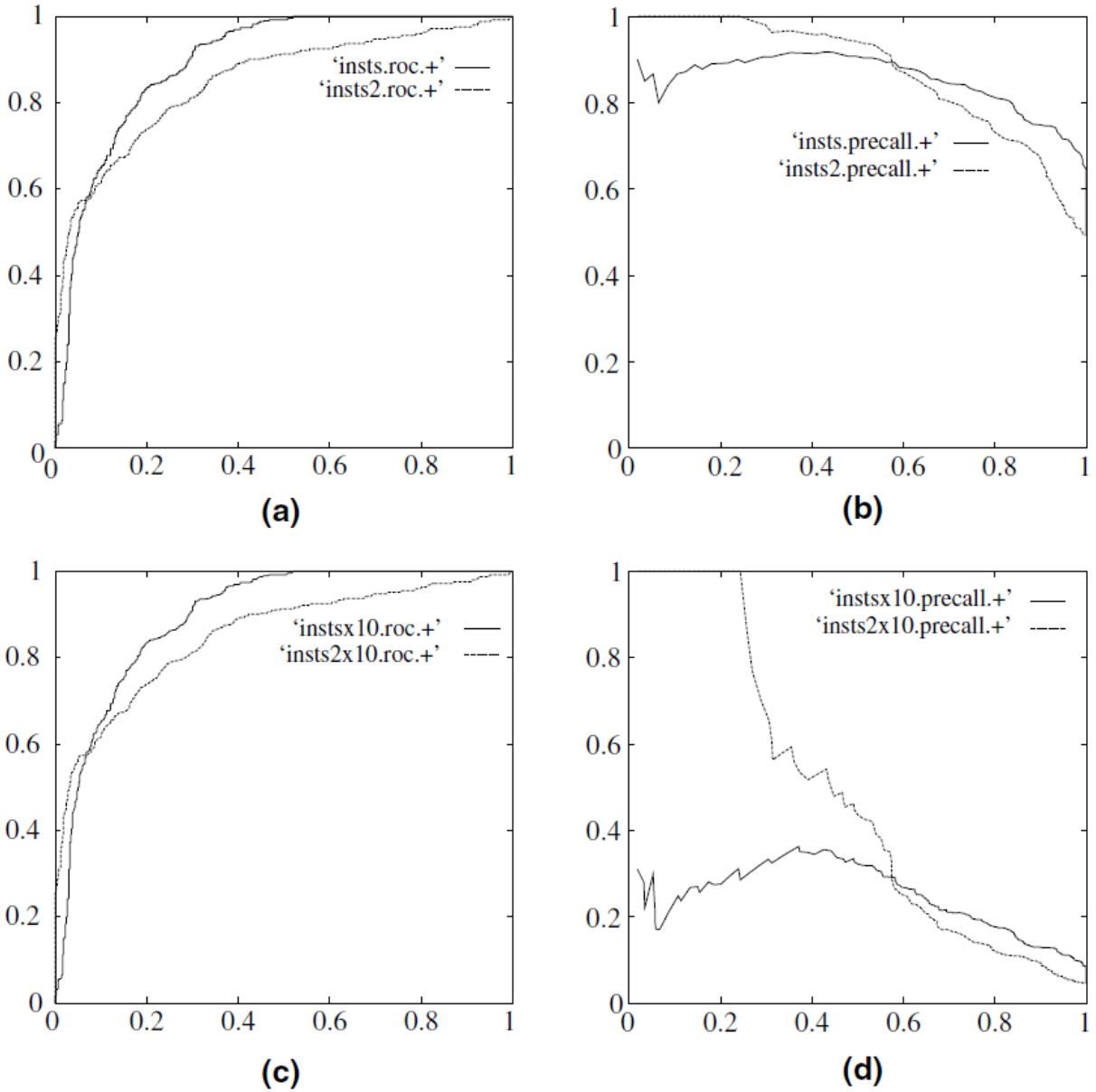


Fig. 5. ROC and precision-recall curves under class skew. (a) ROC curves, 1:1; (b) precision-recall curves, 1:1; (c) ROC curves, 1:10 and (d) precision-recall curves, 1:10.

Relationship between ROC & PR Curves

- One-to-One correspondence between the two curves
- If a curve dominates in ROC space then it dominates in PR space.
- If a curve dominates in PR space then it dominates in ROC space.
- What will be the PR curve for a random classifier?
- What part of an ROC curve impacts the PR curve more?

Reading

- Recommended
  - Davis, Jesse, and Mark Goadrich. 2006. “The Relationship Between Precision-Recall and ROC Curves.” In *Proceedings of the 23rd International Conference on Machine Learning*, 233–40. ICML ’06. New York, NY, USA: ACM. doi:10.1145/1143844.1143874.
  - Fawcett, Tom. 2006. “An Introduction to ROC Analysis.” *Pattern Recogn. Lett.* 27 (8): 861–74. doi:10.1016/j.patrec.2005.10.010.

- Required
  - Alpaydin 2010, Section 19.7

ROC and PR Curves in Scikit-Learn

- http://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_curve.html
- http://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html#example-model-selection-plot-roc-py

- **from sklearn.metrics import ***
- P,R = precision_recall_curve(Y,Z)
- AUCPR = average_precision_score(Y,Z)
- roc_curve, auc

More Scikit Metrics

- http://scikit-learn.org/stable/modules/model_evaluation.html
- F-measure
- Mathews Correlation Coefficient
- Confusion Matrix
- Multiclass metrics

- Read them when you need them!

Measurement of Generalization Performance

- Typically we do not have access to real world test examples
- Use the given "training" set for approximating the generalization performance
- Guidelines
  - There should be "enough" training examples left
  - Test labels should not be used, directly or indirectly, during training
    - Test data (without labels) can be used
  - You should be clear about the intended use and application of the system
  - You should be clear about the objective of performance evaluation

Cross-Validation: K-fold

- Measurement of Generalization Performance
- For estimation of variation
- Divide the data into K folds
  - For k = 1…K
    - Train on K-1 sets leaving the k^{th} set out for validation
    - Validate on the k^{th} set and obtain the performance metrics
  - Report the average and the variation in the performance

Cross-Validation

- If K = Number of examples then this extreme case is called Leave One Out CV (LOOCV)
  - Useful if the amount of data is small
- Stratification
  - Make sure that each fold contains the same number of examples as the overall data
    - If a class has 20 percent examples in the whole dataset, in all samples drawn from the dataset, it should also have approximately 20 percent examples.
- What will the impact on approximated performance with increase in K?

Bootstrapping

- More overlap between samples
- Useful for very small datasets

- For i = 1…b
  - Pick N examples at random from the data set of N examples with replacement
  - Train the classifier on these examples
  - Evaluate the classifier on the original data set and obtain the performance metric

- Average the performance metric to obtain "resubstitution accuracy": acc_s

Bootstrapping

- However, the resubstitution accuracy is an overestimate of the true accuracy due to the inclusion of the training examples in testing
- Let's calculate the chances of a training example in a bootstrap sample not to be in the test sample
  - Given N example, the probability of selecting an example is:
    - 1/N
  - Thus the probability that it is not selected is:
    - 1-1/N
  - Thus, the probability that it is not selected in any of the N picks is:
    - $(1\text{-}1/N)^N = e^{-1} = 0.368$

$$\frac{1}{e} = \lim_{n\to\infty}\left(1-\frac{1}{n}\right)^n.$$

.632 Bootstrap

- Based on this idea, the .632 Bootstrap was proposed by Efron and Tibshirna (1993)
- For i = 1…b
  - Pick N examples at random from the data set of N examples with replacement
  - Train the classifier on these examples
  - Evaluate the classifier on the remaining examples to obtain test error aprroximate $\in_i$
- Evaluate the classifier on the original data set and obtain the resubstitution accuracy acc_s
- The final accuracy estimate is:

$$\mathrm{acc}_{\mathrm{boot}} = \frac{1}{b}\sum_{i=1}^{b}(0.632 \cdot \epsilon 0_i + .368 \cdot \mathrm{acc}_s)$$

Kohavi, Ron. 1995. "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection." In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, 1137–43. IJCAI'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. http://dl.acm.org/citation.cfm?id=1643031.1643047.

Bootstrapping

- There is a .632+ variant which adjusts for "highly overfit rules" such as the nearest-neighbors (or SVMs with RBF kernels with high gammas)
  - Efron, Bradley, and Robert Tibshirani. 1997. "Improvements on Cross-Validation: The .632+ Bootstrap Method." *Journal of the American Statistical Association* 92 (438): 548–60. doi:10.2307/2965703.
- A variant is jackknifing
  - In cross-validation you compute a statistic on the left-out sample(s) using a model built on the kept samples.
  - In jackknifing, you compute a statistic from the kept samples only.

So what to use?

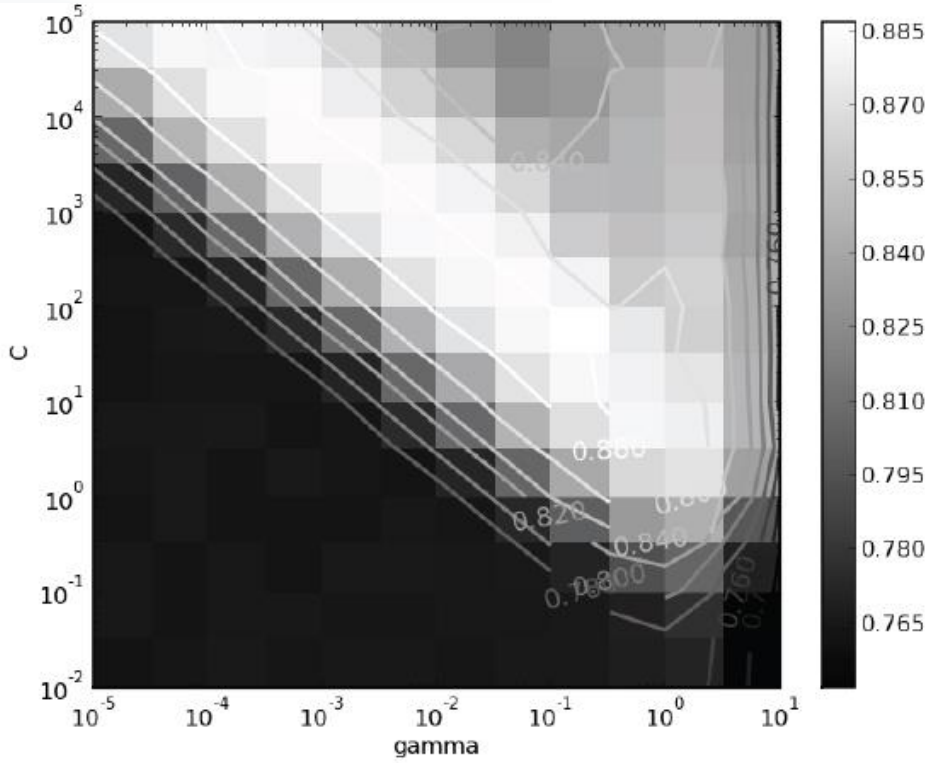
- 10 Fold Cross-Validation is good
- However, for small sample sizes, it can have a large variance in which case you can use LOOCV or the .632 or the .632+ bootstrap

- SCIKIT-LEARN
  - http://scikit-learn.org/stable/model_selection.html

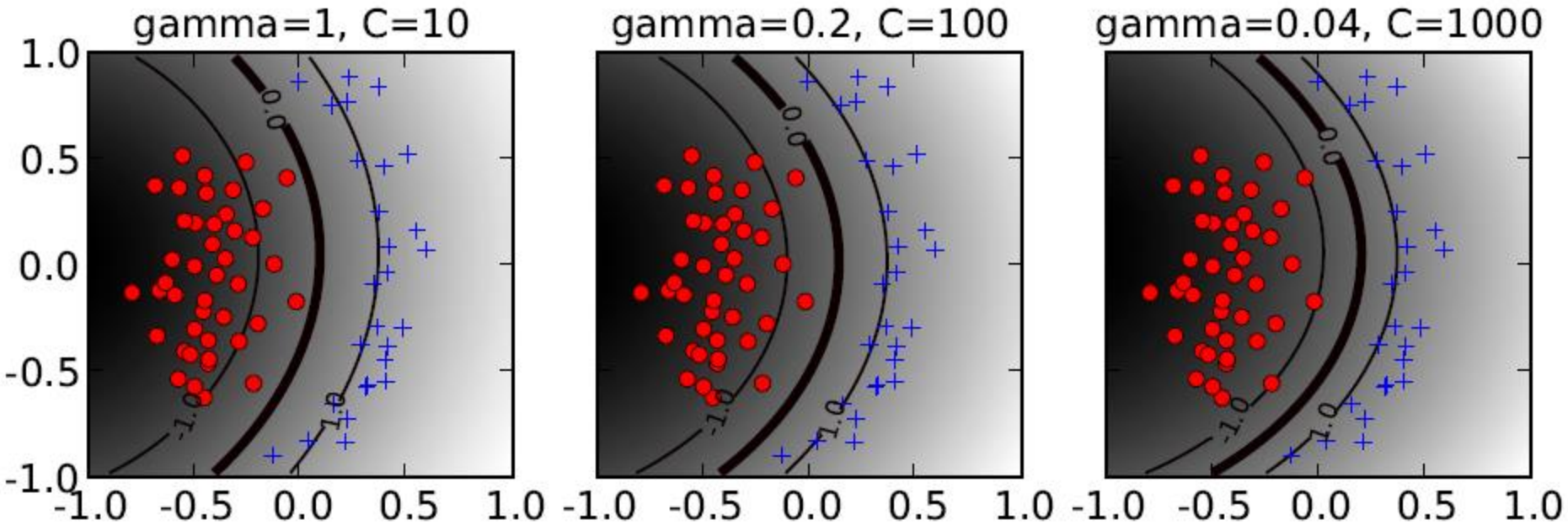
Kohavi, Ron. 1995. "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection." In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, 1137–43. IJCAI'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. http://dl.acm.org/citation.cfm?id=1643031.1643047.

Model Hyperparameter Selection

- Grid Search
  - Exhaustive Search through Cross-Validation
    - Recommended: Nested Cross-Validation or separate test set
- There can be a range of parameter values that yield optimal values and these equivalent points in the parameter space fall along a ridge



Ben-Hur, Asa, and Jason Weston. 2010. "A User's Guide to Support Vector Machines." In *Data Mining Techniques for the Life Sciences*, edited by Oliviero Carugo and Frank Eisenhaber, 223–39. Methods in Molecular Biology 609. Humana Press. http://dx.doi.org/10.1007/978-1-60327-241-4_13.


gamma=1, C=10
gamma=0.2, C=100
gamma=0.04, C=1000
1.0
0.5
0.0
-0.5
-1.0
-1.0 -0.5 0.0 0.5 1.0
0.0
-1.0

Searching for optimal parameters

- Regularization Path Finding
- Gradient Based Approaches
- Evolutionary approaches

- Grid Search in Scikit-learn
  - http://scikit-learn.org/stable/modules/grid_search.html

“Other” ways of selecting parameters

- Selecting gamma
  - Visualize the spread
- Ensuring robustness to parameter changes

Some papers

- Chapter 19 “Design and Analysis of Machine Learning Experiments” Alpaydin, Ethem. 2010. *Introduction to Machine Learning*. Cambridge, Mass.: MIT Press.

- Demšar, Janez. 2006. “Statistical Comparisons of Classifiers over Multiple Data Sets.” *J. Mach. Learn. Res.* 7 (December): 1–30.
- Salvador Garcí, and Francisco Herrera. 2008. “An Extension on ‘Statistical Comparisons of Classifiers over Multiple Data Sets’ for All Pairwise Comparisons.” *Journal of Machine Learning Research* 9 (Dec.): 2677–94.
- Chapelle, Olivier, Vladimir Vapnik, Olivier Bousquet, and Sayan Mukherjee. 2002. “Choosing Multiple Parameters for Support Vector Machines.” *Machine Learning* 46 (1-3): 131–59. doi:10.1023/A:1012450327387.
- Hastie, Trevor, Saharon Rosset, Robert Tibshirani, and Ji Zhu. 2004. “The Entire Regularization Path for the Support Vector Machine.” *J. Mach. Learn. Res.* 5 (December): 1391–1415.

- Fernández-Delgado, Manuel, Eva Cernadas, Senén Barro, and Dinani Amorim. 2014. “Do We Need Hundreds of Classifiers to Solve Real World Classification Problems?” *Journal of Machine Learning Research* 15: 3133–81
- Forman, George, and Ira Cohen. 2004. “Learning from Little: Comparison of Classifiers Given Little Training.” In *Knowledge Discovery in Databases: PKDD 2004*, edited by Jean-François Boulicaut, Floriana Esposito, Fosca Giannotti, and Dino Pedreschi, 161–72. Lecture Notes in Computer Science 3202. Springer Berlin Heidelberg. http://link.springer.com/chapter/10.1007/978-3-540-30116-5_17.
- Salperwyck, C., and V. Lemaire. 2011. “Learning with Few Examples: An Empirical Study on Leading Classifiers.” In *The 2011 International Joint Conference on Neural Networks (IJCNN)*, 1010–19. doi:10.1109/IJCNN.2011.6033333.

CASE STUDY

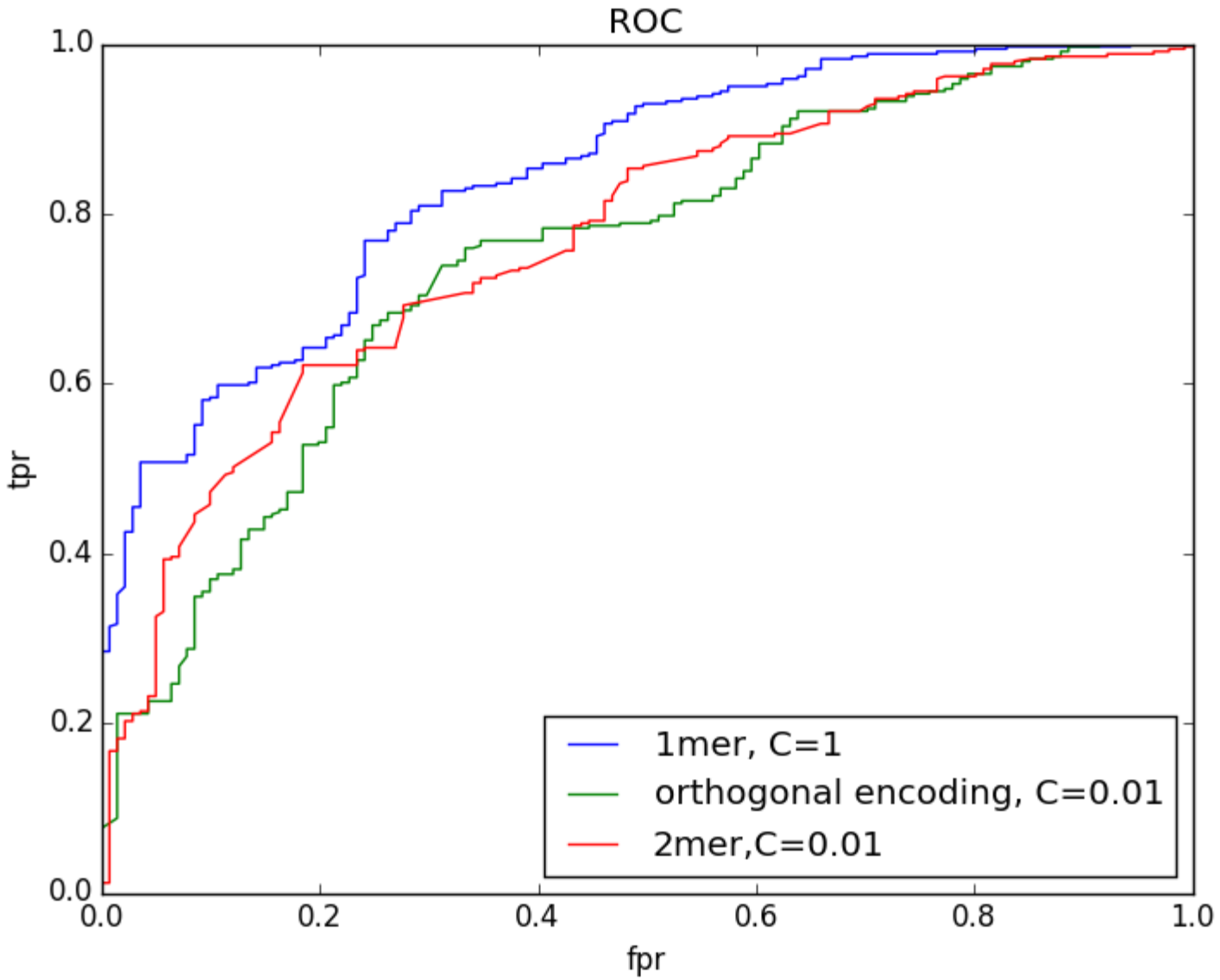
- Amyloid Prediction by Farzeen
- Training Data
- Test Data

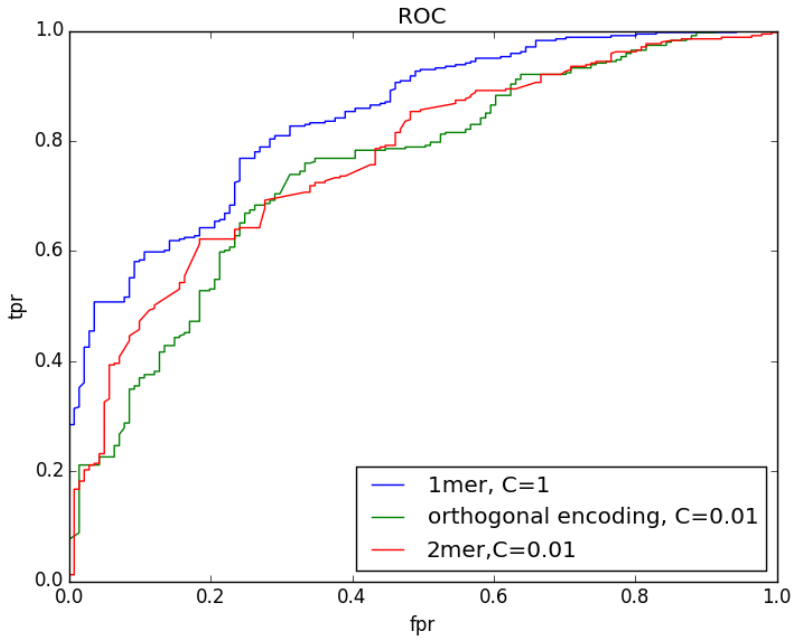
Selecting “C”

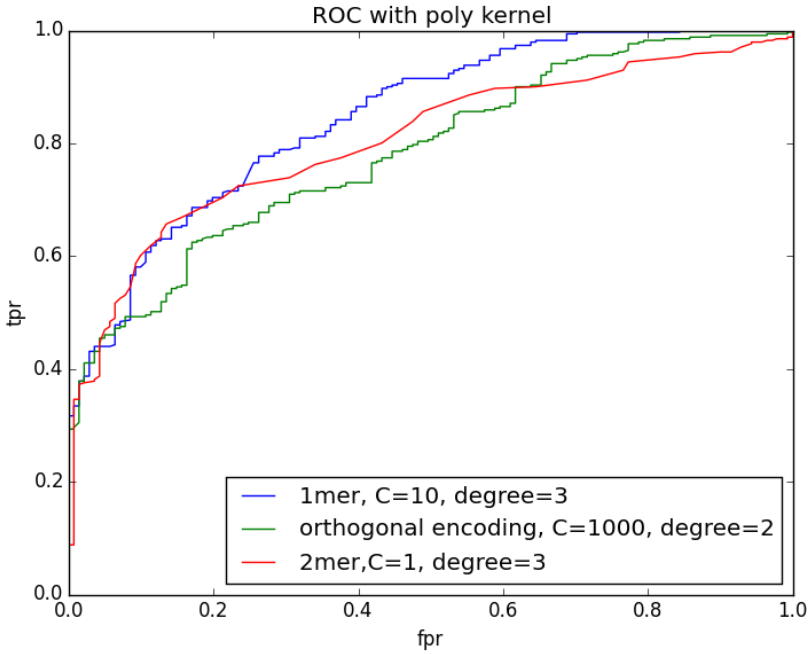
C, 1 mer	1st fold of training	2nd fold of training	3rd fold of training	average
0.0001	0.78	0.65	0.822	0.75
0.001	0.78	0.67	0.824	0.75
0.01	0.78	0.67	0.823	0.757
0.1	0.79	0.705	0.87	0.788
1	0.801	0.75	0.84	**0.797**
10	0.809	0.76	0.809	0.7926
100	0.809	0.76	0.809	0.792

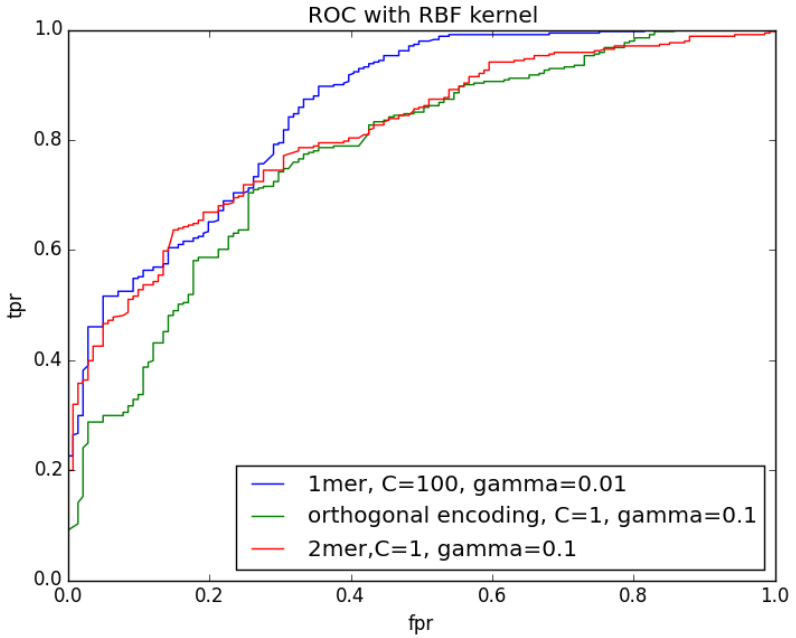
Grid Search

C	gamma=0.001	gamma=0.01	gamma=0.1	gamma=1	gamma=10	gamma=100	gamma=1000
0.0001	0.734	0.733	0.734	0.731	0.525	0.525	0.525
0.001	0.734	0.734	0.734	0.734	0.55	0.546	0.546
0.01	0.734	0.732	0.735	0.737	0.553	0.546	0.546
0.1	0.736	0.7329	0.735	0.704	0.553	0.546	0.546
1	0.724	0.733	**0.758**	0.745	0.597	0.546	0.546
10	0.7340	0.750	0.725	0.751	0.597	0.546	0.546
100	0.746	0.687	0.725	0.751	0.597	0.546	0.546
1000	0.746	0.687	0.725	0.751	0.597	0.546	0.546


ROC
tpr
fpr
0.0
0.2
0.4
0.6
0.8
1.0
1mer, C=1
orthogonal encoding, C=0.01
2mer,C=0.01


ROC
tpr
fpr
1mer, C=1
orthogonal encoding, C=0.01
2mer,C=0.01


ROC with poly kernel
tpr
fpr
1mer, C=10, degree=3
orthogonal encoding, C=1000, degree=2
2mer,C=1, degree=3


ROC with RBF kernel
tpr
fpr
1mer, C=100, gamma=0.01
orthogonal encoding, C=1, gamma=0.1
2mer,C=1, gamma=0.1

End of Lecture

if you can't measure it, you can't manage it.