AI-503 Advanced Machine Learning

Announcement

Choose a day feasible for all the students in case we have to arrange a make-up class.

- Tuesday and Thursday before lunch
- Any time on Wednesday

We may have to reschedule a few classes in the month of March.

Last time

Concluded SVMs

Regularization in SVMs

Weight vector perspective

- What would cause a bigger change in the output for a small change in the input?
- Small value of $||\boldsymbol{w}||^2$ or large?
- Large values of C result in smaller margin
  - => Large value of $||\boldsymbol{w}||^2$
  - => More change in decision score for small change in input
  - => Less regularization
- Lowering the value of C increases regularization

Complexity perspective

Feature Scaling

SVMs and the methods based on Euclidean distance (e.g. KNN) are sensitive to feature scales.

Rescale your data before applying these methods.

You may observe a lot of difference in performance before and after applying scaling

StandardScaler (sklearn)

- 0 mean 1 variance

MinMaxScaler (sklearn)

- Min 0 max 1

Read pg 132-140 of Introduction to Machine Learning with Python

Today

How to compare ML models

- My model is better than yours

How can you select optimal hyperparameters?

- C, gamma, kernel, K…

Hyper-parameter Tuning and Performance Evaluation of ML models

Two important tasks

- We need to choose the best possible hyperparameters
  - Hyperparameters over which the model is expected to give the best **performance**
- We want to know how good our method would **perform** on real test data
  - And also how good it is as compared to its counterparts

The catch?

- We don't have real test data (most of the times)

Also,

- What do we mean by **performance?**

Classification Performance Metrics

Something that can quantify how good a classifier is.

Supposing you have the test data with its true labels and a trained ML model, how would you judge how good your model is?

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

Accuracy

How accurate is the classifier?

(No. of correct predictions)/ (Total No. of predictions)

Accuracy

Consider a classifier that calls everything an apple.

Now test it over a dataset consisting of 90 apples and 10 oranges.

What would be the accuracy?

Classification Performance

- A classifier (or any machine learning model) can be viewed as a function $y = f(x|\theta)$ 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%**	

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

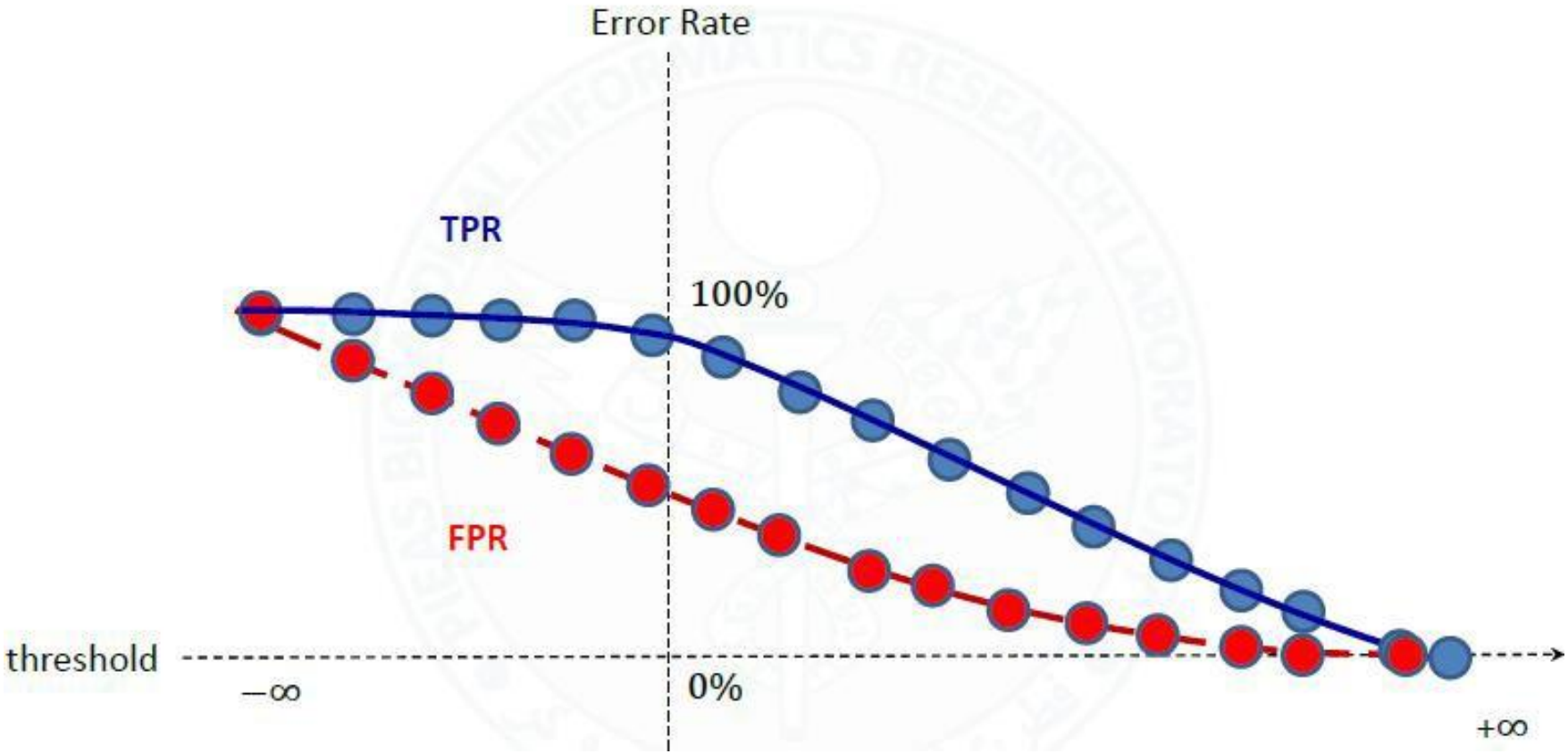
Confusion matrix

		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−) $= \frac{\text{FNR}}{\text{TNR}}$	

Behavior of metrics

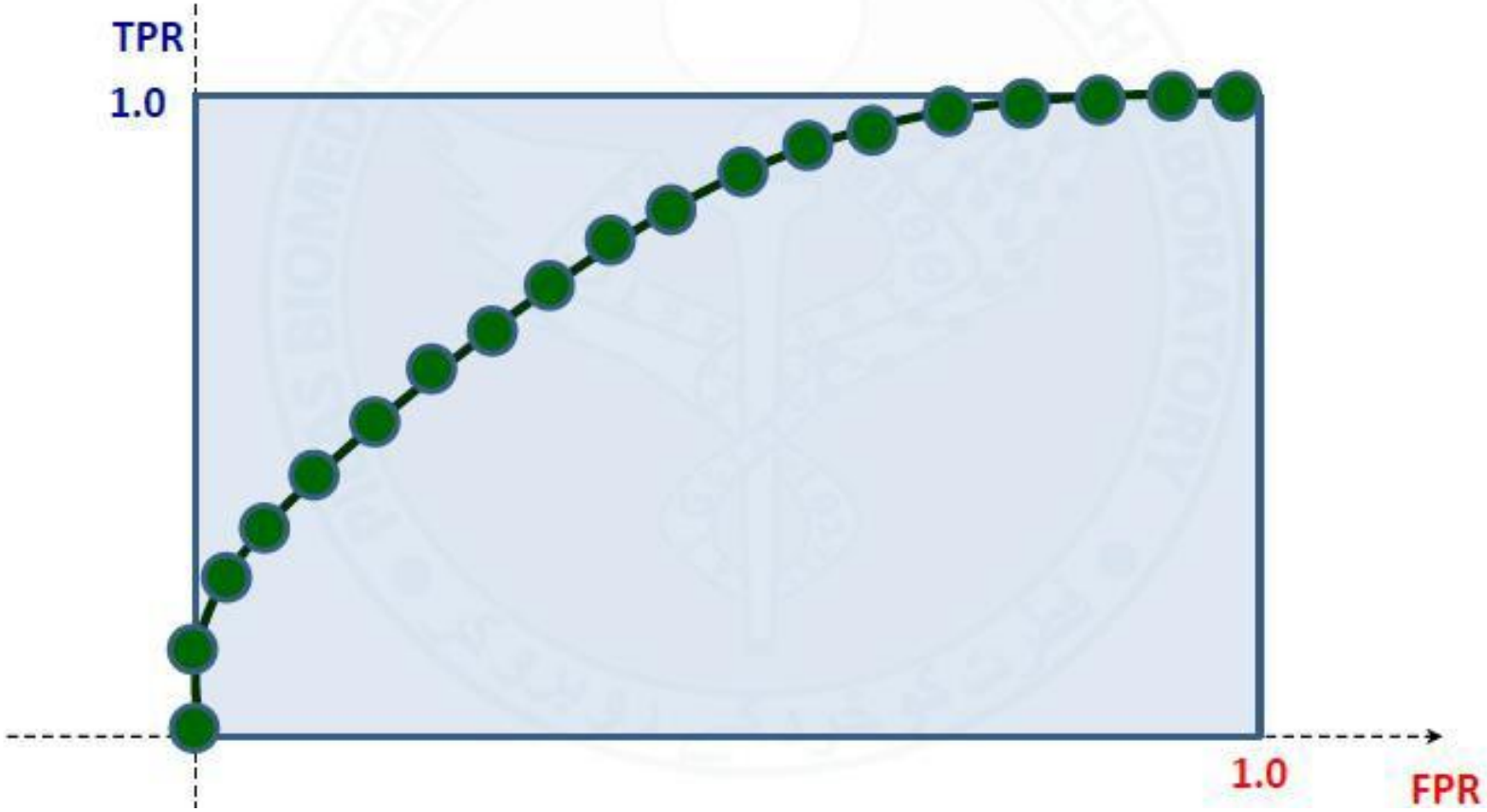
- What will be the behavior of TPR with increase in threshold of the classifier?
- How will FPR behave?
- How will Precision behave?
- Can TPR decrease with increase in threshold?

FPR vs. TPR Curve



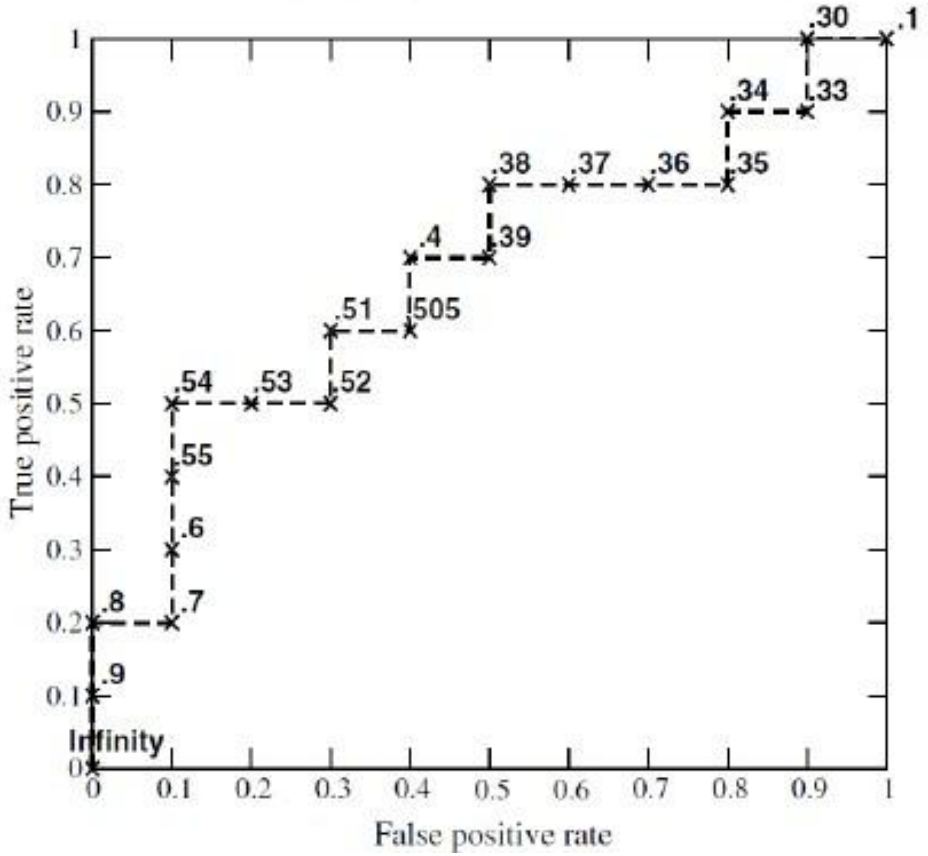
Receiver Operating Characteristics Curve

- 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

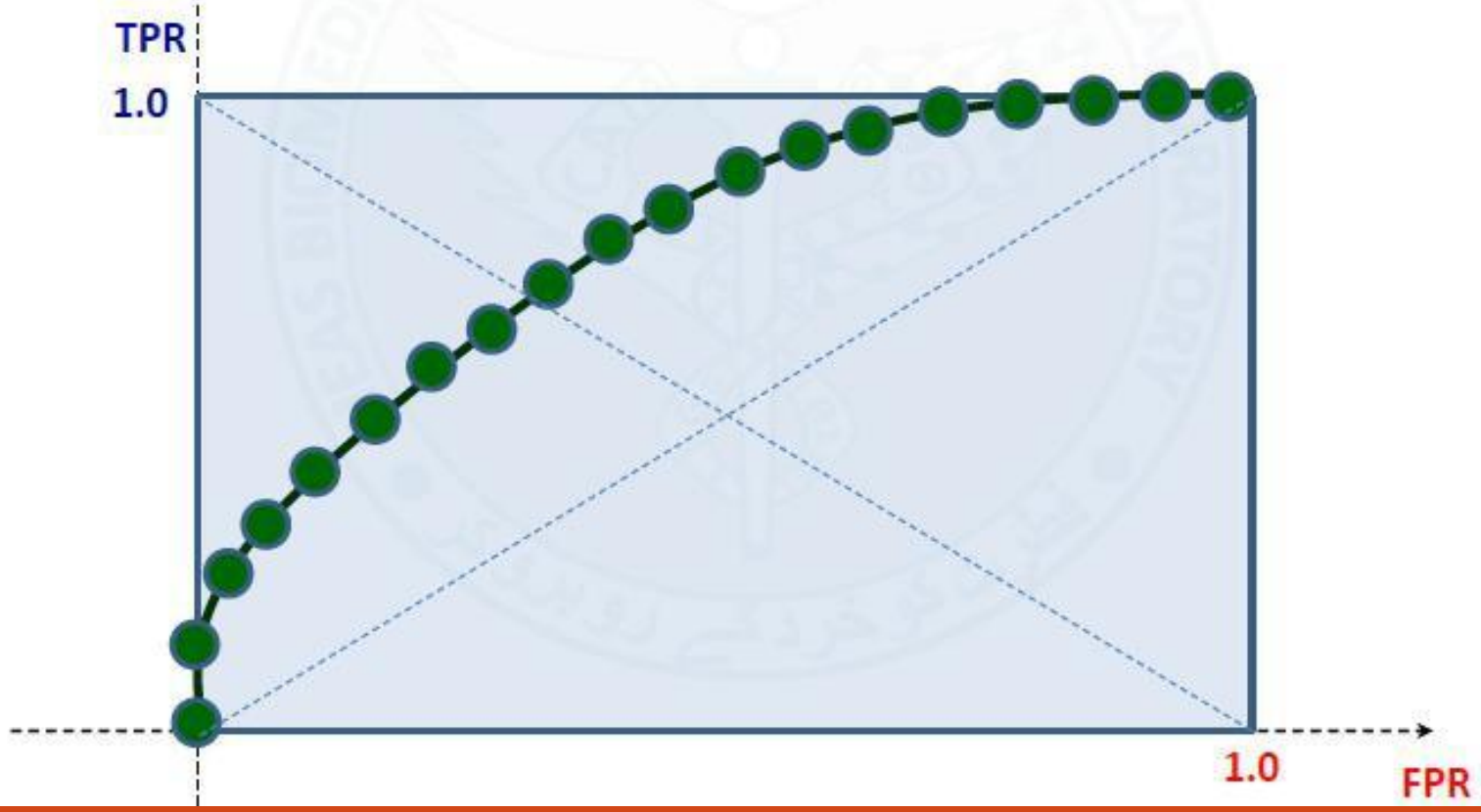


ROC curves

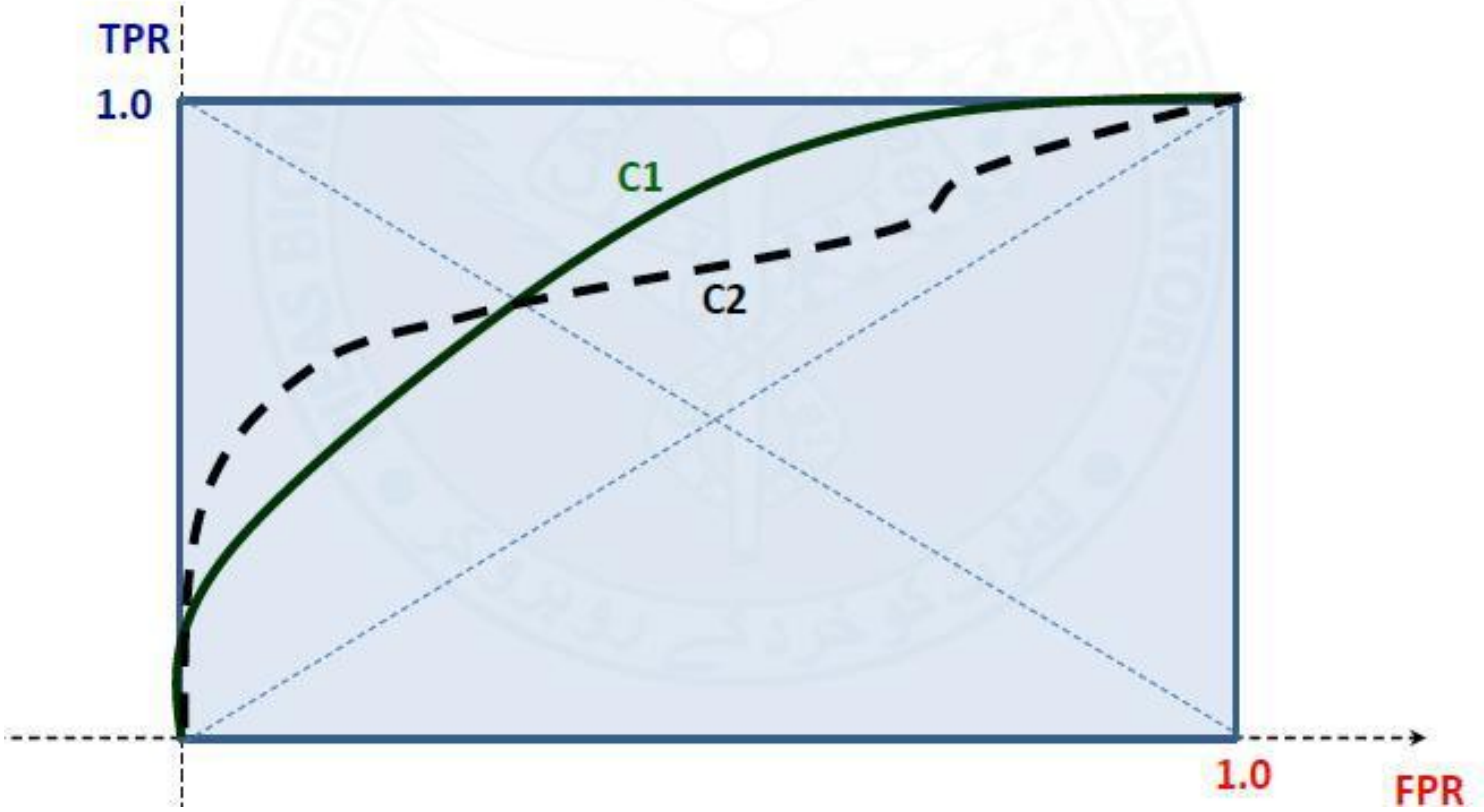
- What will be ROC curve for a perfect classifier?
- What will the ROC Curve of a random classifier look like?
- What will the ROC curve of a classifier that always predicts the positive class look like?
- What are the underlying assumptions of the ROC curve?
- What part of the ROC curve is the most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

AUC-ROC

- The area under the ROC curve is a quality metric

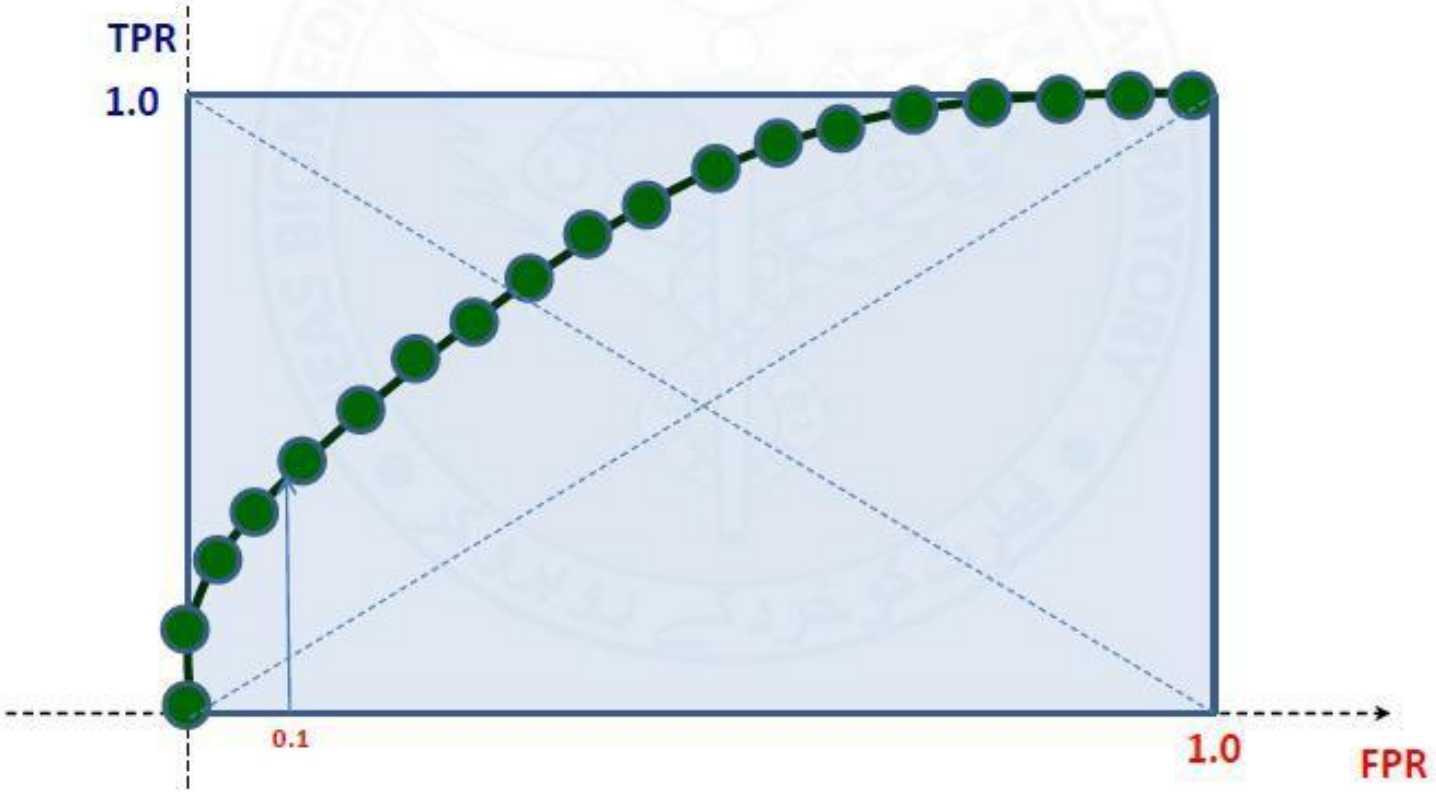


Which one is better?



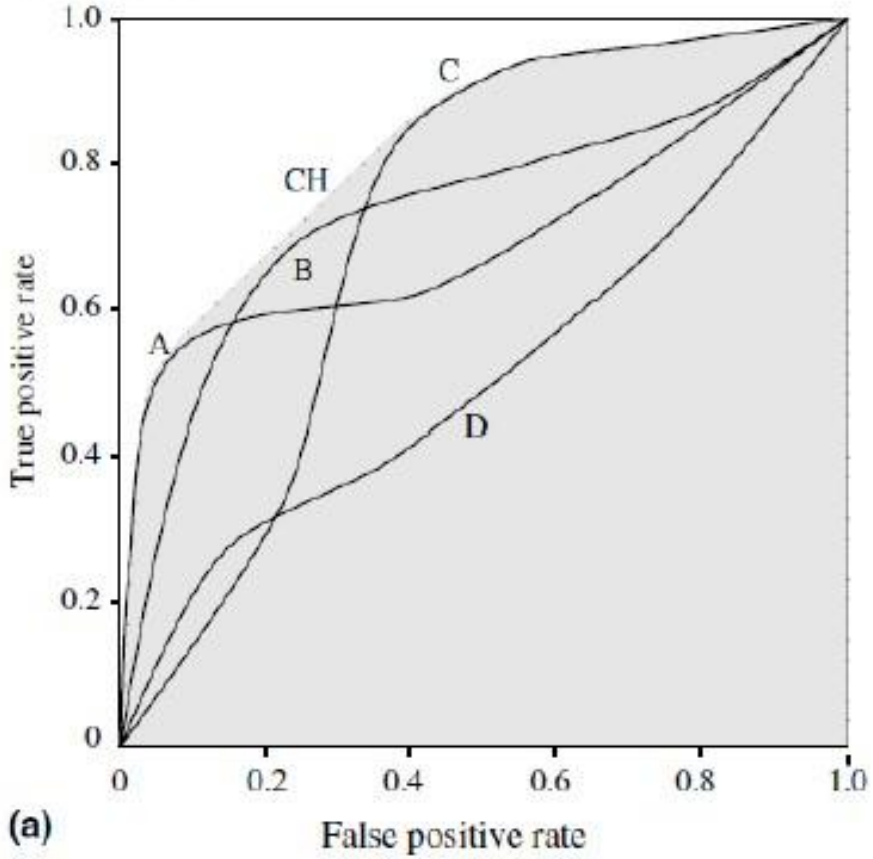
AUC ROC-N

- 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



Multi-class ROC Curves

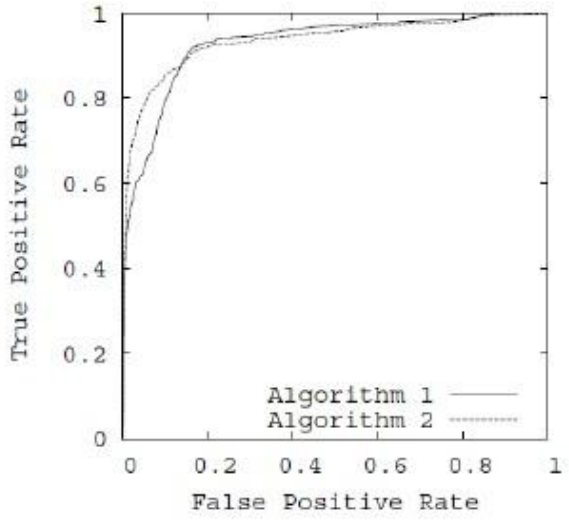
- Can also make multiple class ROC curves
  - One vs. Rest
- AUC-ROC can also be computed
  - Pairwise

Properties

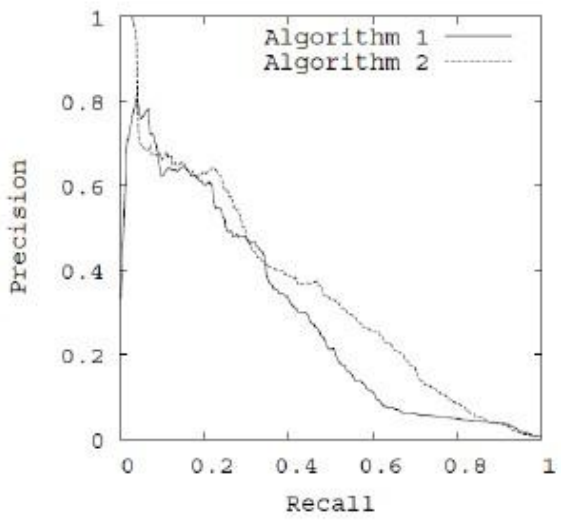
- Class Imbalance?
- When to Use?
- What to focus on?
  - FPR?
  - TPR?

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

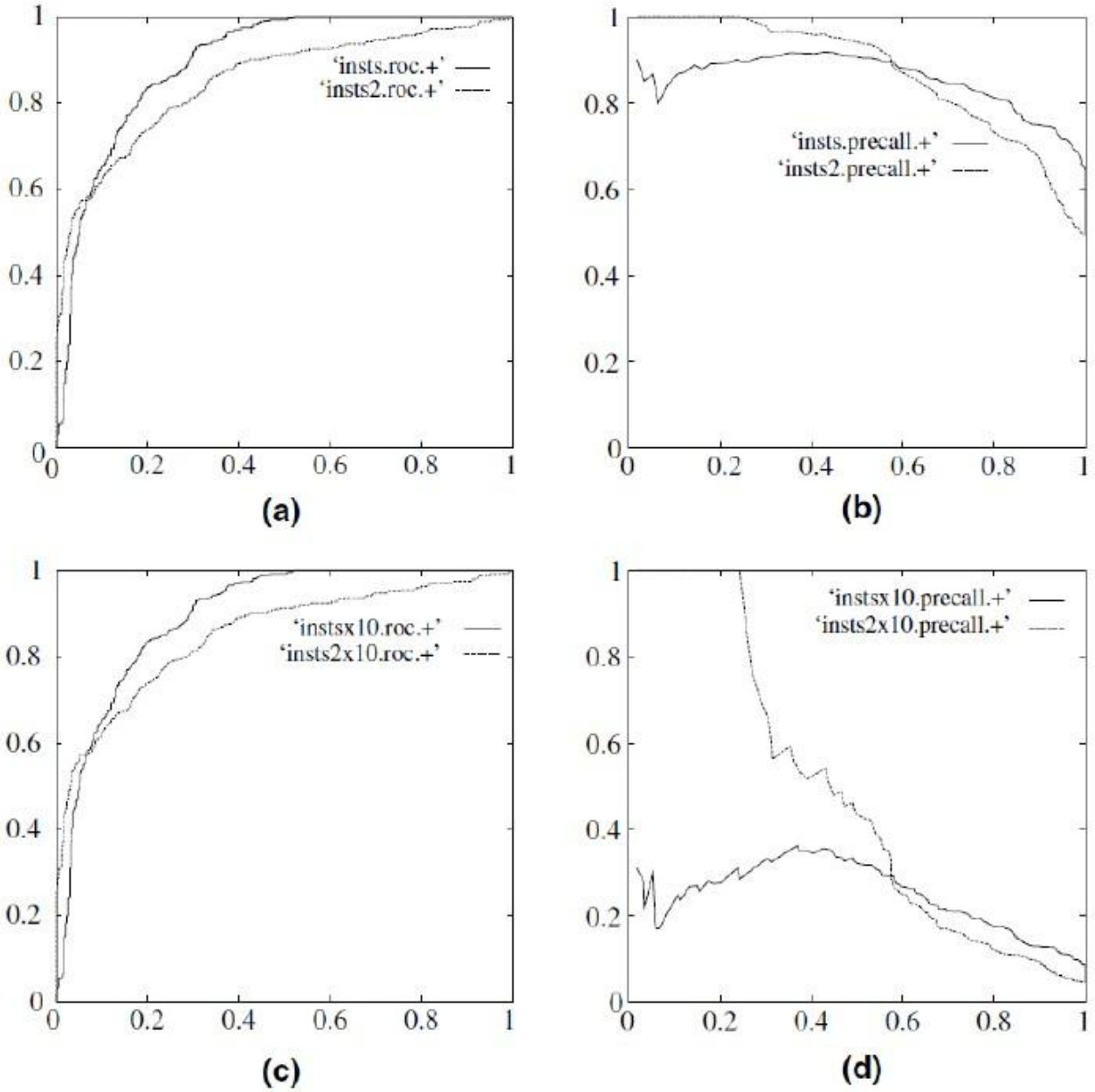
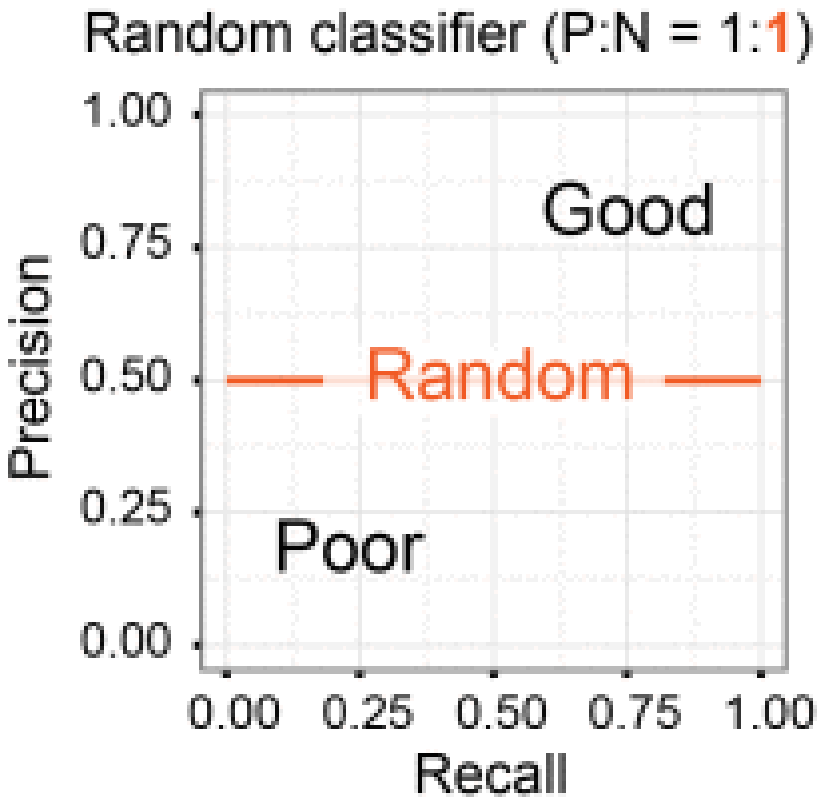
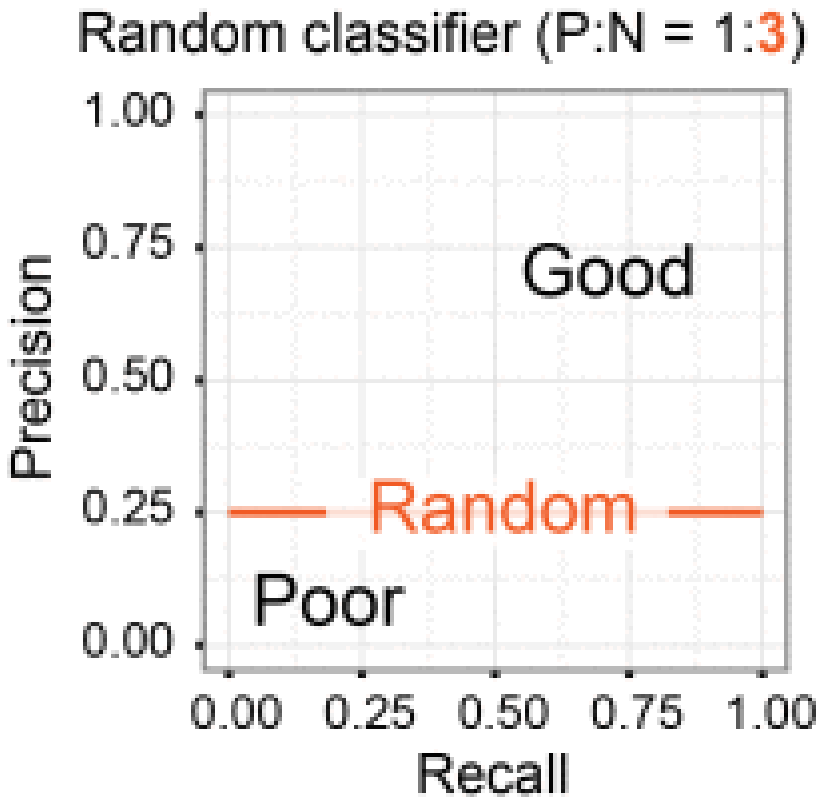


Fig. 5. ROC and precision-recall curves under class skew. (a) ROC curves, 1:1; (b) precision-recall curves, 1:1; (c) ROC curves, 1:10 and (d) precision-recall curves, 1:10.

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?


Random classifier (P:N = 1:1)
1.00
0.75
0.50
0.25
0.00
Precision
Good
Random
Poor
0.00
0.25
0.50
0.75
1.00
Recall


Random classifier (P:N = 1:3)
1.00
0.75
0.50
0.25
0.00
Precision
Good
Random
Poor
0.00
0.25
0.50
0.75
1.00
Recall

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

Demo. . .

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!

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

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

Issues

- The variance of our estimate increases as the size of the test set decreases.
- A small increase in the pessimistic bias when we decrease the size of the training set

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 be 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

So what to use?

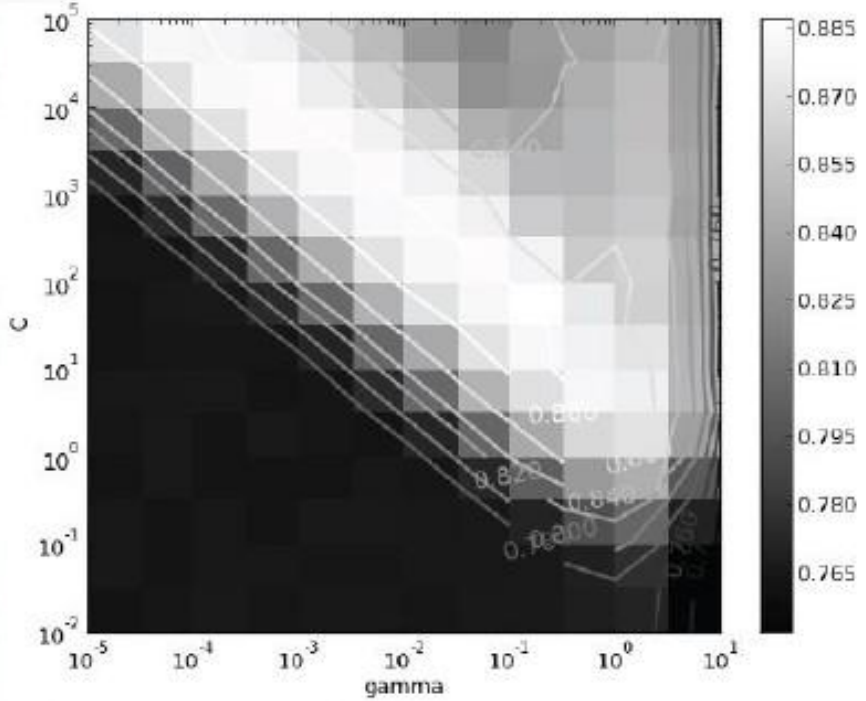
- 10 Fold Cross-Validation is good
- However, for small sample sizes, it can have a large variance in which case you can use LOOCV or the .632 or the .632+ bootstrap

- 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.

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

Parameter Selection

- http://hyperopt.github.io/
- http://hyperopt.github.io/hyperopt-sklearn/
- https://automl.github.io/auto-sklearn/stable/api.html
- https://en.wikipedia.org/wiki/Xgboost
- http://www.kdnuggets.com/2017/01/current-state-automated-machine-learning.html

"Other" ways of selecting parameters

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