Final Results of Machine Learning Tasks

STAT639

John M. Niehaus
Department of Political Science
Texas A&M University

Unsupervised Learning

Summary:

4 clustering algorithms, all on various PC subspaces.

1. K-means
2. GMM
3. Hierarchical
4. Density-based

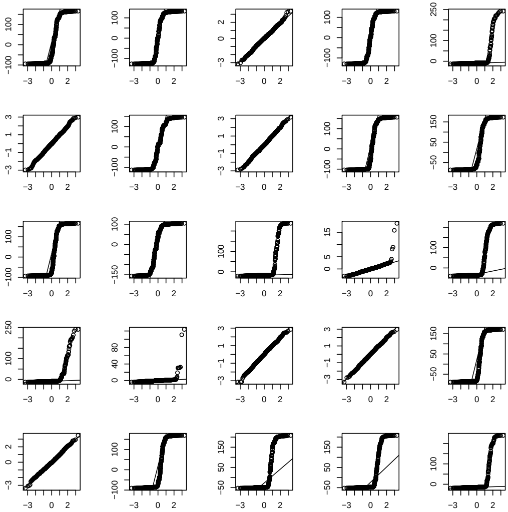
Notes:

1. Normality assumptions likely not met
2. Euclidean distance in high dimension breaks down

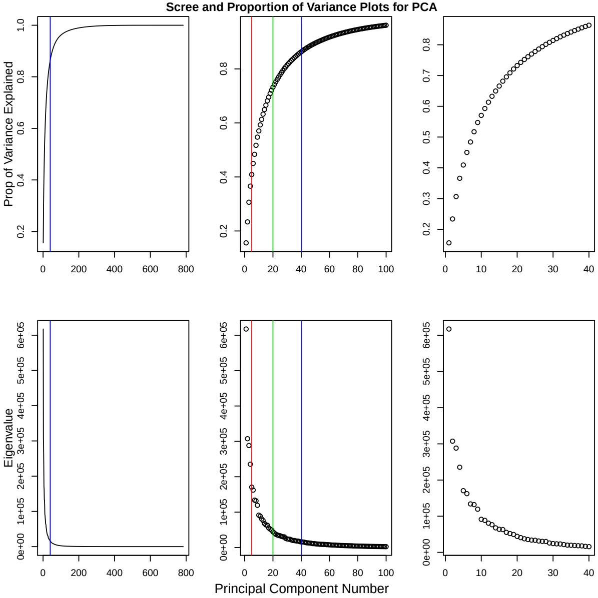
Unsupervised: Normality Assumptions

A necessary but insufficient condition:

$$\boldsymbol{y}_j \sim N(\mu_j, \sigma_j) \quad \forall \quad j \in (1, 2, \ldots, p)$$



Unsupervised: Dimension Reduction



Unsupervised: Choosing K

1. CH Index:

$$CH(K) = \frac{BSS/K - 1}{TWSS/n - k}$$

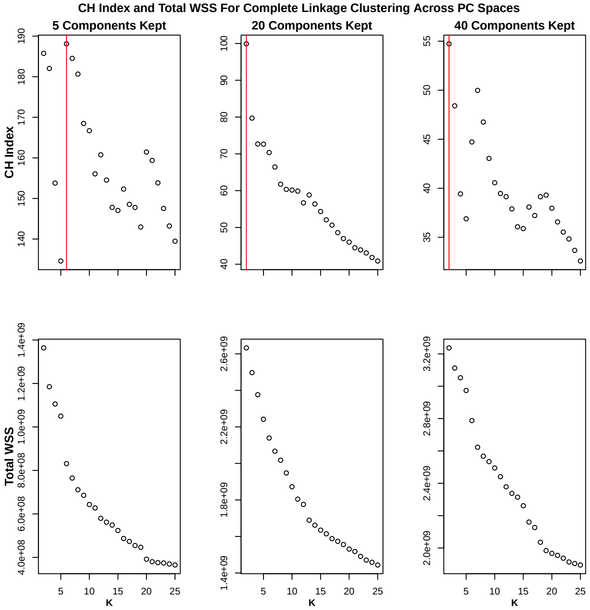
   Choose $\arg\max_K CH(K)$ to

   - Maximize between cluster distance
   - Minimize within cluster distance

2. Use BIC for GMM due to being fit by MLE

(Caliński and Harabasz 1974; See also: Charrad et al. 2014 for 30 other indices for choosing K)

Unsupervised: Complete Linkage



Unsupervised: DBSCAN Results

	Eps	Min. Points	N Clusters
5 PCs	595	40	2
20 PCs	1255	40	2
40 PCs	1455	25	3

Table: CH Maximizing Parameters to DBSCAN

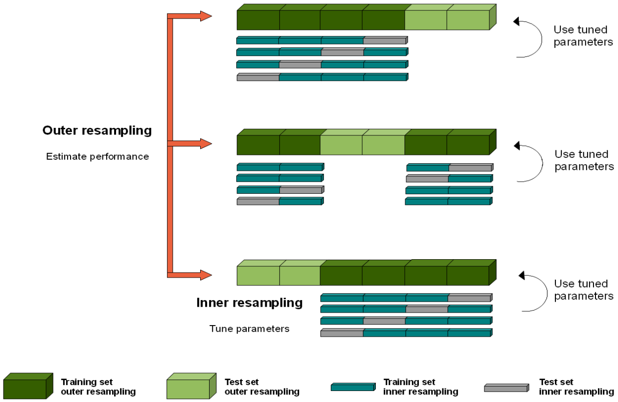
Supervised Learning Task

Summary:

- Considered 6 algorithms
  1. KNN
  2. Naive Bayes
  3. Logistic Elastic Net
  4. SVM
  5. Random Forest
  6. Stochastic Gradient Boosting

Supervised: Repeat, Nested CV

Nested CV:



Repeat this process 5 times, getting 5 CV estimates of the risk. Average these estimates.

Image source: https://weina.me/nested-cross-validation/

Supervised: Stochastic Boosting

Stochastic boosting wins all inner CV loops by roughly 10 percentage points.

What is it?

- Stochastic – Allows random subsamples of observations and features to be considered in each boosting round (like RF)
- Requires some minimal loss reduction to honor a split. Thus, depth of tree is an upper bound, not a certainty.
- The rest is the same as the boosting we've learned about

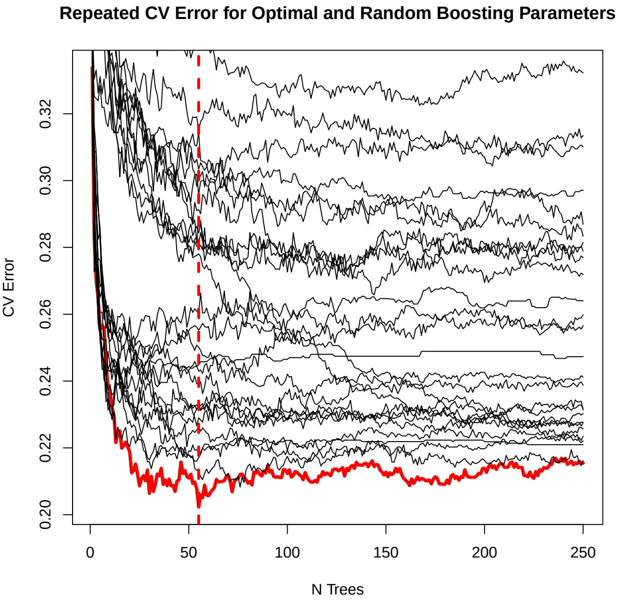
(See Friedman 2002, for more on stochastic boosting.)

Supervised: Test Error

	λ	Depth	Col. Sample	γ	Trees	CV
Outer Fold 1	0.0001	9	0.25	0.01	137	0.1993
Outer Fold 2	0.01	9	0.5	0	44	0.1972
Outer Fold 3	0.01	7	0.25	0	71	0.1841
Outer Fold 4	0.001	9	0.25	0.5	32	0.2055
Outer Fold 5	0.01	7	0.25	1	24	0.2009
Outer Fold 6	0.01	7	0.5	1	28	0.1668

Table: Optimal Inner Nested CV Errors for One Repitition

Supervised: Optimal Parameters



Supervised: Conclusion

Optimal tuning parameters are:

1. Boosting
2. 55 trees
3. Learning rate .0001
4. Max depth of 9
5. feature sampling ratio of 0.25
6. minimum loss reduction of 0.5

which has an estimated test error rate of 21.6%.

References I

Caliński, Tadeusz and Jerzy Harabasz. 1974. "A dendrite method for cluster analysis." *Communications in Statistics-theory and Methods* 3(1):1–27.

Charrad, Malika, Nadia Ghazzali, Veronique Boiteau and Azam Niknafs. 2014. "NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set." *Journal of Statistical Software* .

Friedman, Jerome H. 2002. "Stochastic gradient boosting." *Computational statistics & data analysis* 38(4):367–378.