Lecture 7: Machine Learning of the microbiome

Paul McMurdie

UNSUPERVISED LEARNING: CLUSTERING

How do we understand clustering?

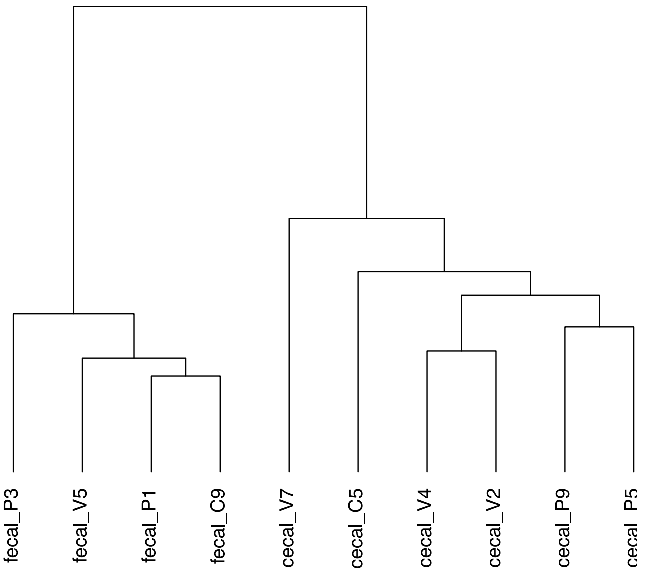
- What does it mean for the data to be clustered?
- What meaning do the clusters have?
- How do you know the data can be clustered?

Definition

- Clustering analysis – methodologies for describing proximity between objects
- Hierarchical clustering – a set of descriptive techniques for grouping objects by similarity
- Discrete clustering – a set of techniques for assessing membership of objects in one of several closely groups.

Hierarchical clustering

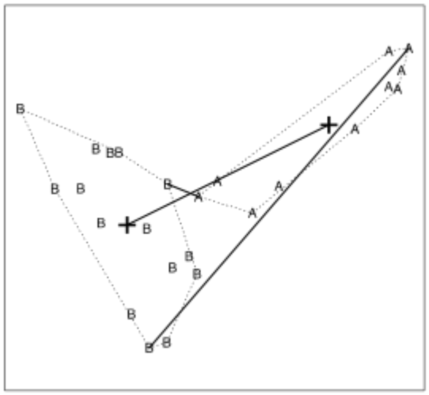
- Organize objects in dendrograms (usually binary);
- Objects that are more similar are closer to each other on the tree;
- For any set of objects one can find a dendrogram! Everything clusters!
- How to tell if the clustering is meaningful?



Hierarchical clustering algorithm

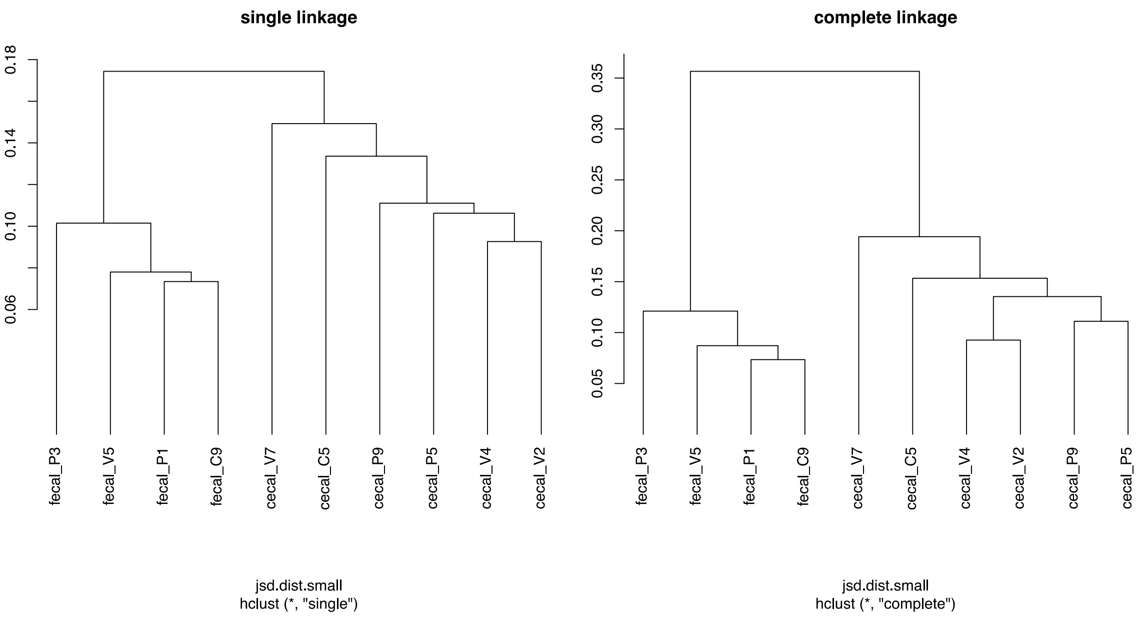
- Start with a dissimilarity matrix
- Join the 2 most closely related objects
- Remove the joined objects from the matrix
- Add a new object that represents the joint group (complete, average, single)
- Repeat until no objects remain in the matrix

Linkage types



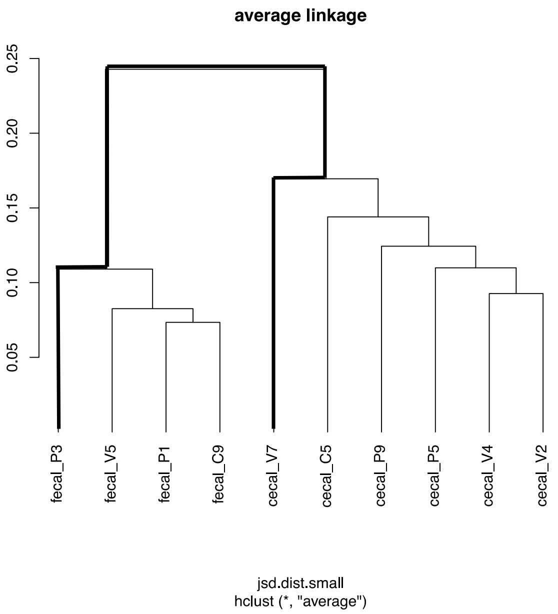
- Complete linkage: distance from the furthest objects apart
- Average linkage: average distance between objects
- Single linkage: distance from the closest objects apart

Hierarchcal clustering example



Cophenetic distance

- Distance induced by the dendrogram is called *cophenetic* distance.
- This distance may be different from the original distance used to construct the dendrogram.

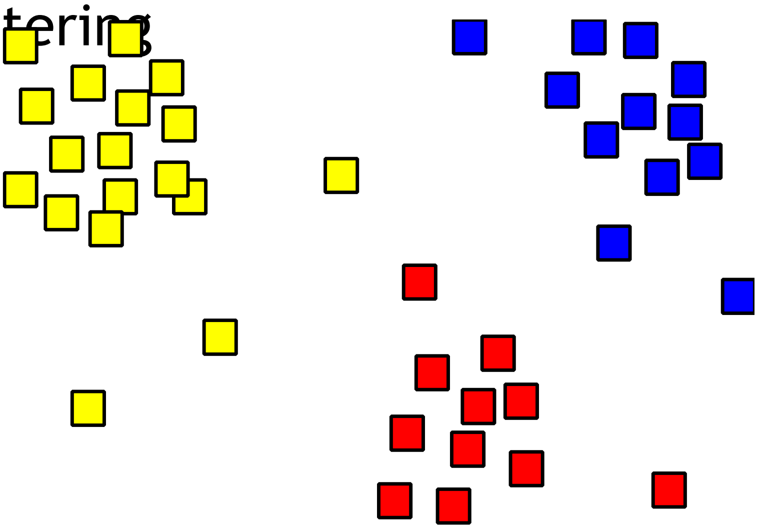


In R

- `hclust` performs hierarchical clustering
- `cophenetic` computes cophenetic distance on the dendrogram

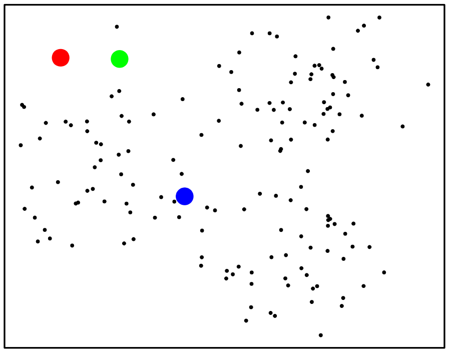
Discrete clustering

- K-means clustering
- PAM (partitioning around medoids) clustering

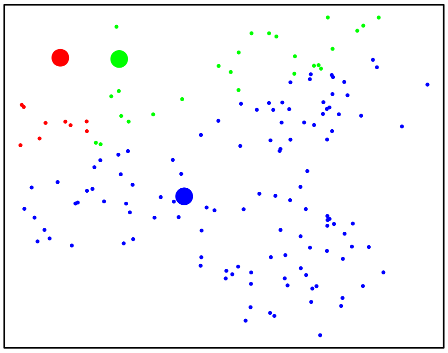


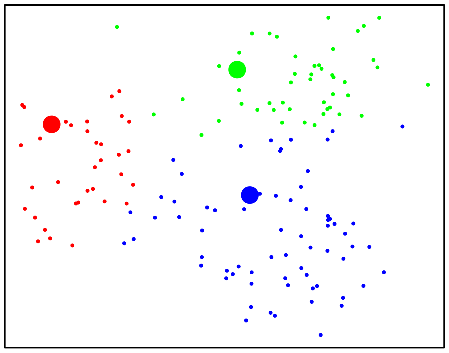
K-means clustering

- **Initialize:** Pick K random points as cluster centers
- **Iterate:**
  - Assign points to closest cluster center
  - Update cluster center location to the mean of the assigned points
- **Stop** when no points change cluster assignment (convergence)

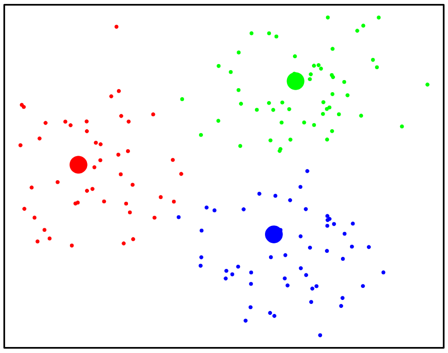


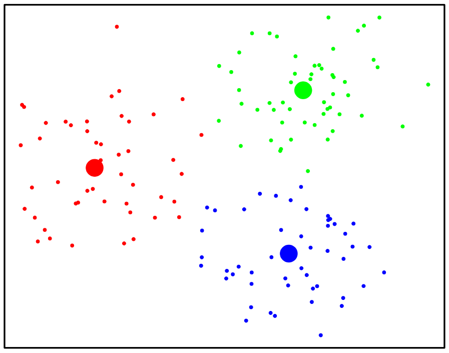
K-means clustering



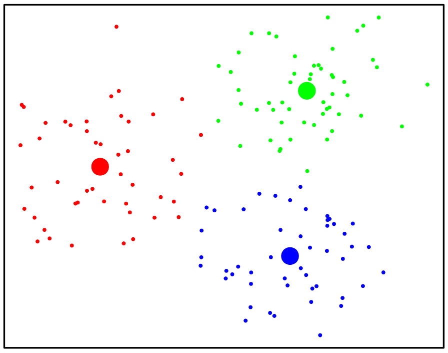


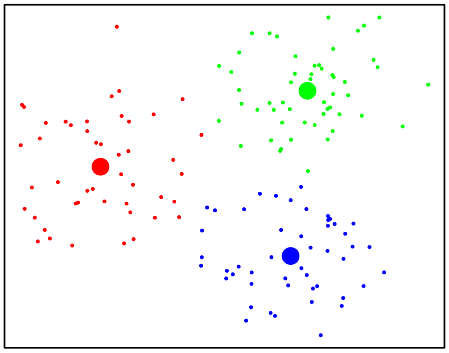
K-means clustering





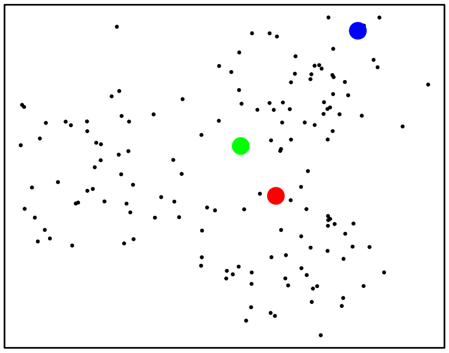
K-means clustering



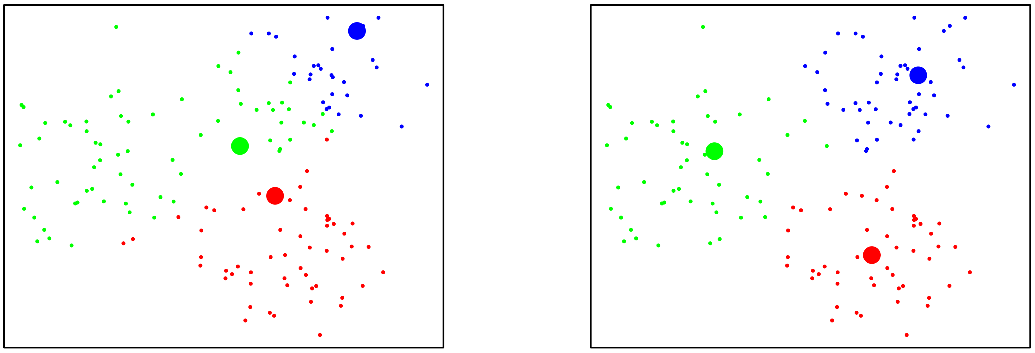


Partitioning around medoids clustering

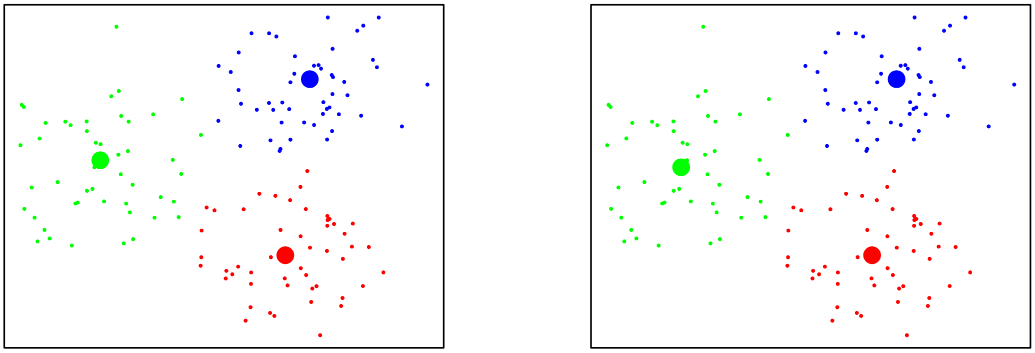
- **Initialize:** Select K of the points to be the centers of the clusters
- **Iterate:**
  - Assign points to the closest cluster center.
  - For each cluster center:
    - Replace center with point that minimizes total distance within the cluster
- **Stop** when no cluster center has changed



K-medoids



K-medoids



In R:

- Libraries: `cluster` and `clusterSim`
- `pam`: partitioning around medoids algorithm
- `clusGap`: gap statistic

How to select the number of clusters?

- Use measures of how good the clusters describe the structure of the data for varying number of clusters.
- F-statistic: Calinski-Harabasz index
- Silhouette method
- Gap statistic: a metric based on within group distances defined using permutations

F-statistic

- Let
  - SSW is the sum of squares within the clusters;
  - SSB is the sum of squares among the clusters.
- F [CH-index]= $(SSB/(K-1))/(SSW/(n-K))$
  - Ratio of average between cluster distance and average within cluster distance
- Larger index value indicates better clustering:
  - When distance between clusters is maximized so is the F index;
  - When within cluster variability is low the index is higher.

Silhouette

- For each point i let:
  - a(i) is average distance to other objects within the same cluster;
  - b(i) distance to the closest object outside the cluster.
- $s(i) = [b(i) – a(i)]/\max(a(i), b(i))$
- $-1 \leq s(i) \leq 1$
- s(i) closer to 1 indicates best clustering; when a(i) is vanishingly small and b(i) is much larger than a(i).

Gap Statistic

$$D_r = \sum_{i,i' \in C_r} d_{ii'} \qquad (1)$$

be the sum of the pairwise distances for all points in cluster r, and set

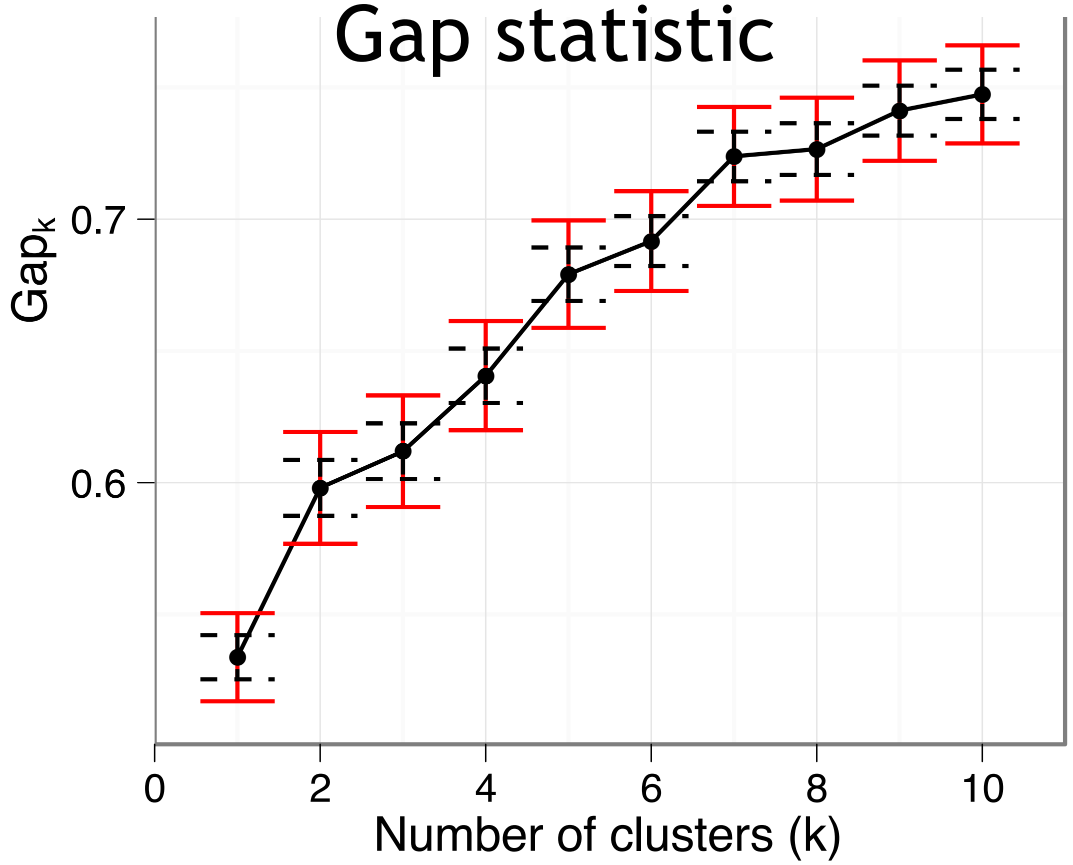
$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r. \qquad (2)$$

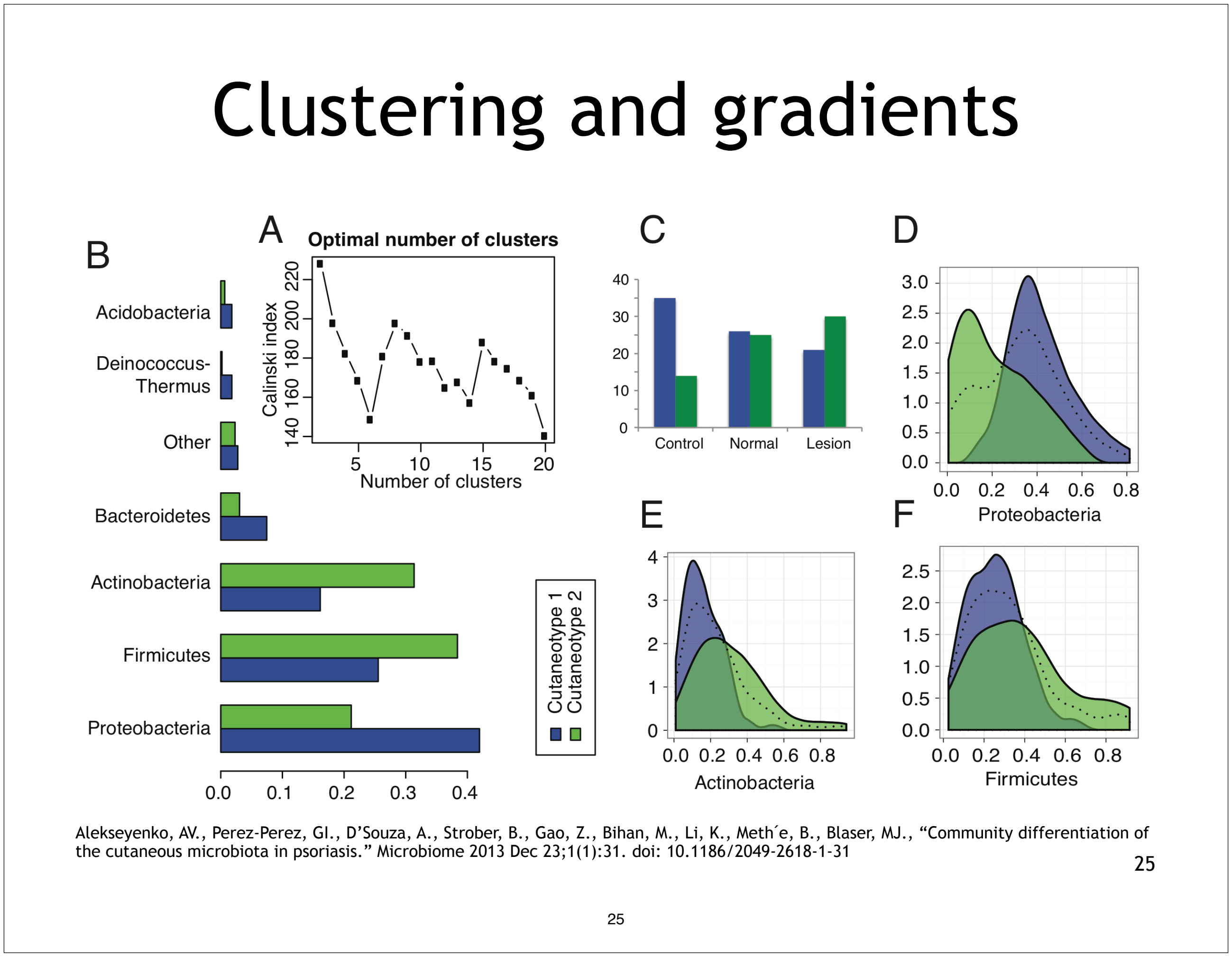
$$\mathrm{Gap}_n(k) = E_n^*\{\log(W_k)\} - \log(W_k), \qquad (3)$$

Computed using “random” clustering.

Larger values of the gap statistic correspond to better clustering.

Gap statistic





Clustering and gradients

Alekseyenko, AV., Perez-Perez, GI., D'Souza, A., Strober, B., Gao, Z., Bihan, M., Li, K., Meth´e, B., Blaser, MJ., "Community differentiation of the cutaneous microbiota in psoriasis." Microbiome 2013 Dec 23;1(1):31. doi: 10.1186/2049-2618-1-31

SUPERVISED LEARNING: CLASSIFICATION

What are the main elements of predictive downstream analysis?

1. **Model selection**

   Out of many possible models find the ones that are most likely to be accurate (and also have other desired properties).

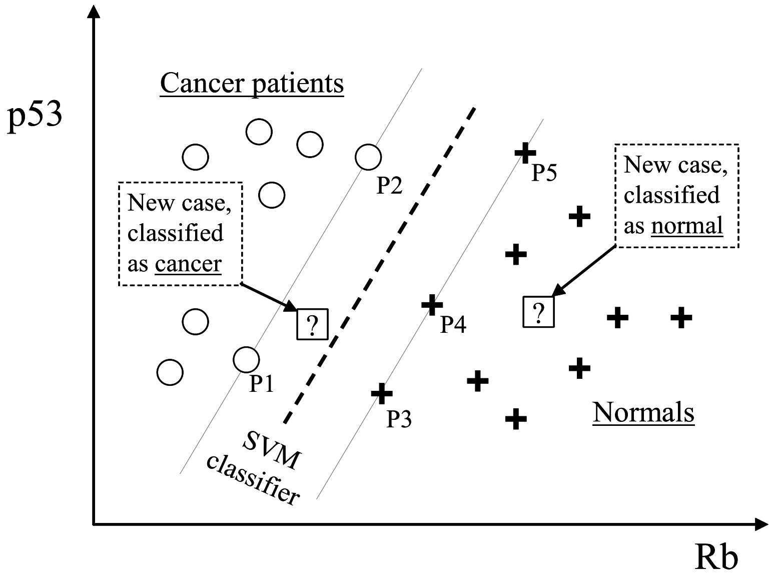
2. **Error estimation**

   Estimate how accurate the final model will be in future applications (i.e., in the population where we sampled from).

Very important Model Selection + Error Estimation method:

Repeated Nested n-Fold Cross Validation (RNCV)

Supervised learning: a geometrical interpretation


p53
Cancer patients
P2
New case, classified as cancer
?
P1
SVM classifier
P5
New case, classified as normal
?
P4
P3
Normals
Rb

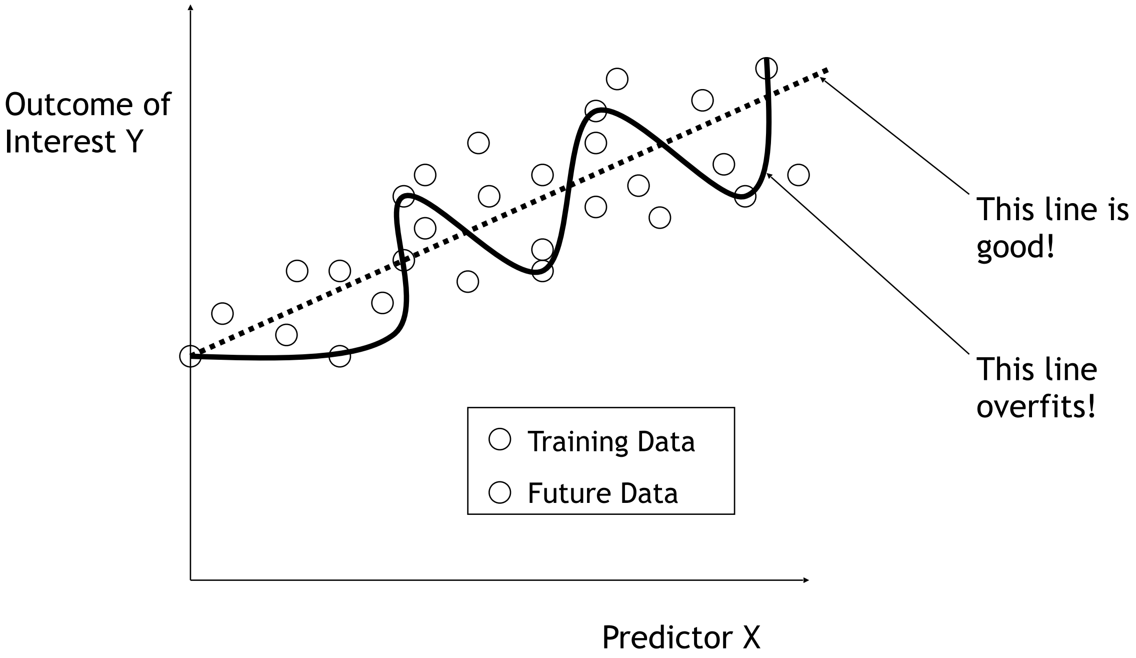
High-dimensionality (especially with small samples) causes:

- Some methods do not run at all (classical multiple regression)
- Some methods give bad results (KNN, Decision trees)
- Very slow analysis
- Very expensive/cumbersome clinical application
- Tends to “overfit”

Two (very real and very unpleasant) problems: Over-fitting & Under-fitting

- **Over-fitting (a model to your data)**= building a model that is good in original data but fails to generalize well to fresh data
- **Under-fitting (a model to your data)**= building a model that is poor in both original data and fresh data

Over/under-fitting are directly related to the complexity of the decision surface and how well the training data is fit



Outcome of Interest Y

Predictor X

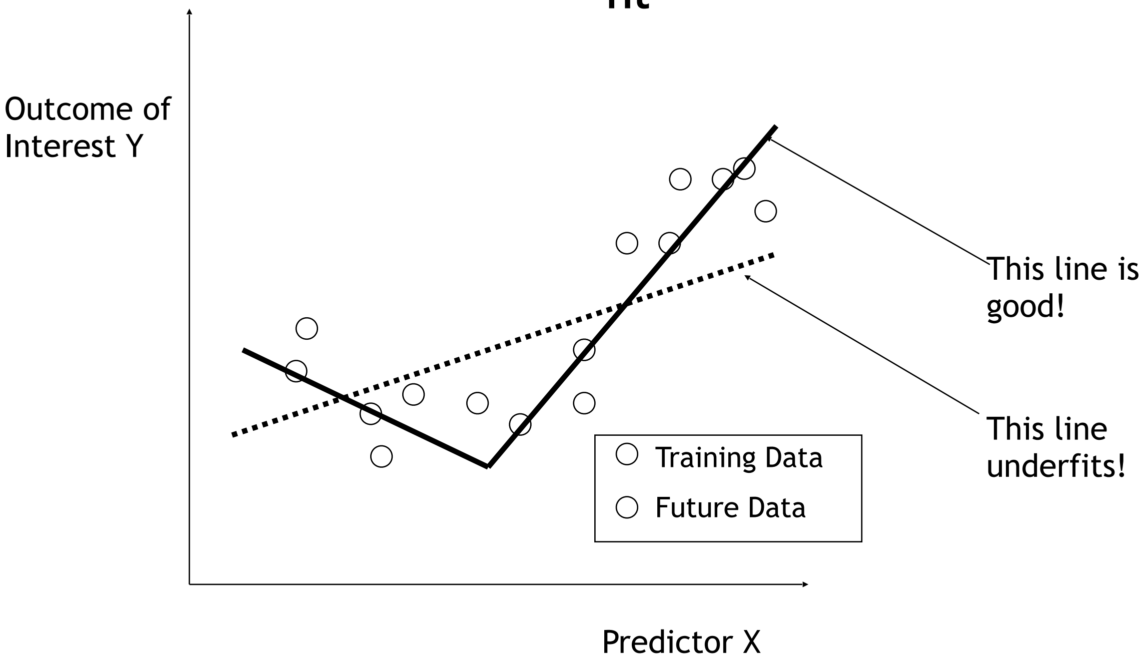
○ Training Data

○ Future Data

This line is good!

This line overfits!

Over/under-fitting are directly related to the complexity of the decision surface and how well the training data is fit



Outcome of Interest Y

Predictor X

○ Training Data

○ Future Data

This line is good!

This line underfits!

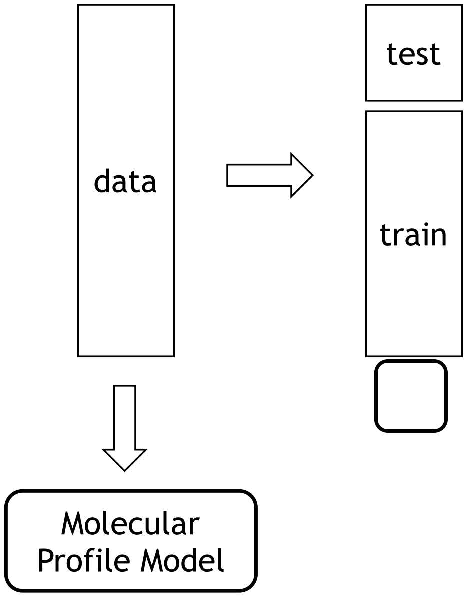
Successful data analysis methods balance training data fit with complexity

- Too complex signature (to fit training data well) ➔overfitting (i.e., signature does not generalize)
- Too simplistic signature (to avoid overfitting) ➔ underfitting (will generalize but the fit to both the training and future data will be low and predictive performance small).

What is overfitting? What is its relationship with high dimensionality?

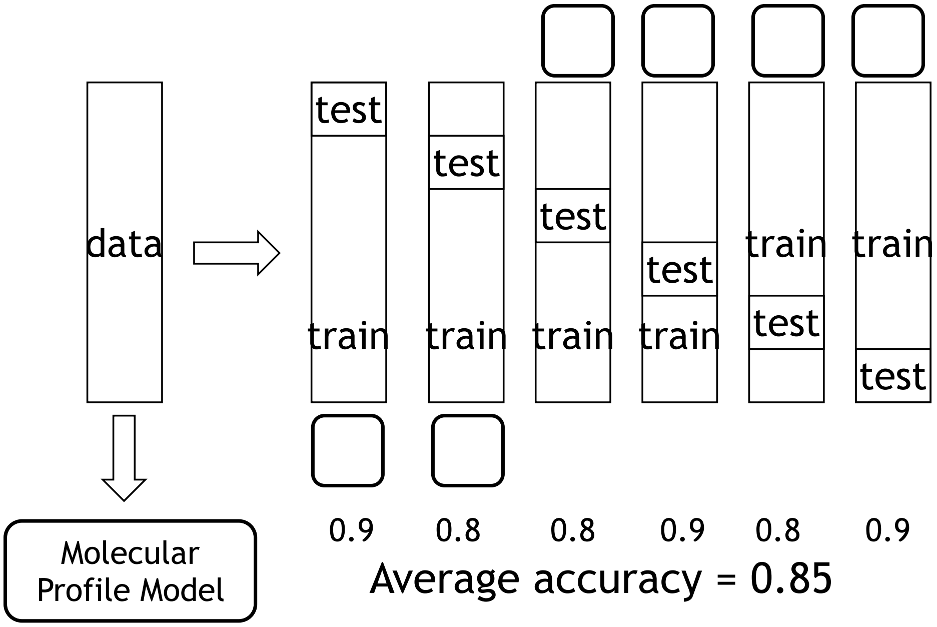
1. **Overfitting:** when we create a model that accurately captures characteristics of our discovery dataset but fails to perform well in the populations where the discovery data was sampled from.
2. All else being equal, high dimensionality makes overfitting easier to occur.

Hold-out validation method



35

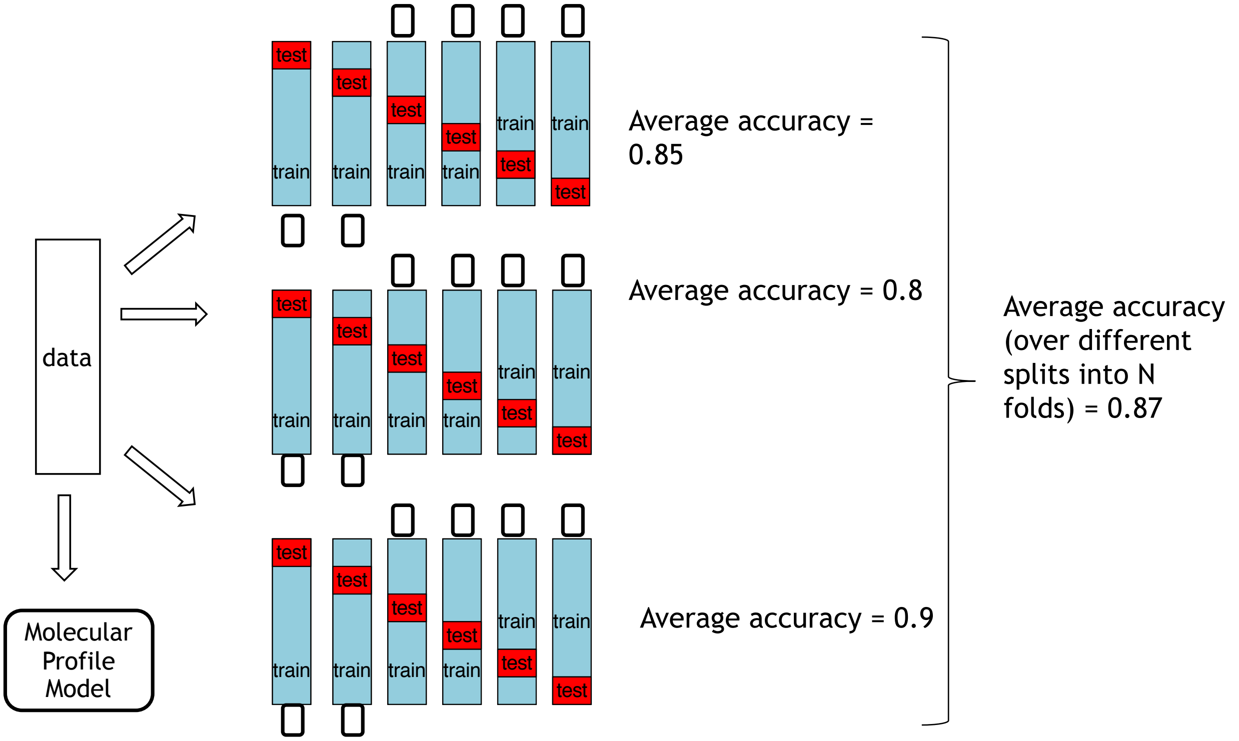
N-Fold Cross-validation



Average accuracy = 0.85

36

Repeated N-Fold Cross-validation



Average accuracy = 0.85

Average accuracy = 0.8

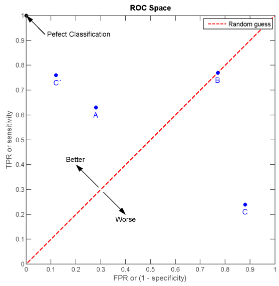
Average accuracy = 0.9

Average accuracy (over different splits into N folds) = 0.87

Measures of classification error

- Accuracy: proportion of correct classifications
  - The number of times the classifier gives the correct result divided by the total number of test cases.
- Area under receiver-operator characteristic curve (AUC).

		Truth		
		Positive	Negative	
Test outcome	Positive	True positive	False positive	**Precision:** #TP/ #PPositives
	Negative	False negative	True negative	**NPP:** #TN/ #PNegative
		Sensitivity: #TP/ #Positives	**Specificity:** #TN/ #Negatives	**Accuracy**



http://en.wikipedia.org/wiki/File:ROC_space-2.png

Comparison of State of the Art Methods for Microbiomic marker + Signature Discovery 1

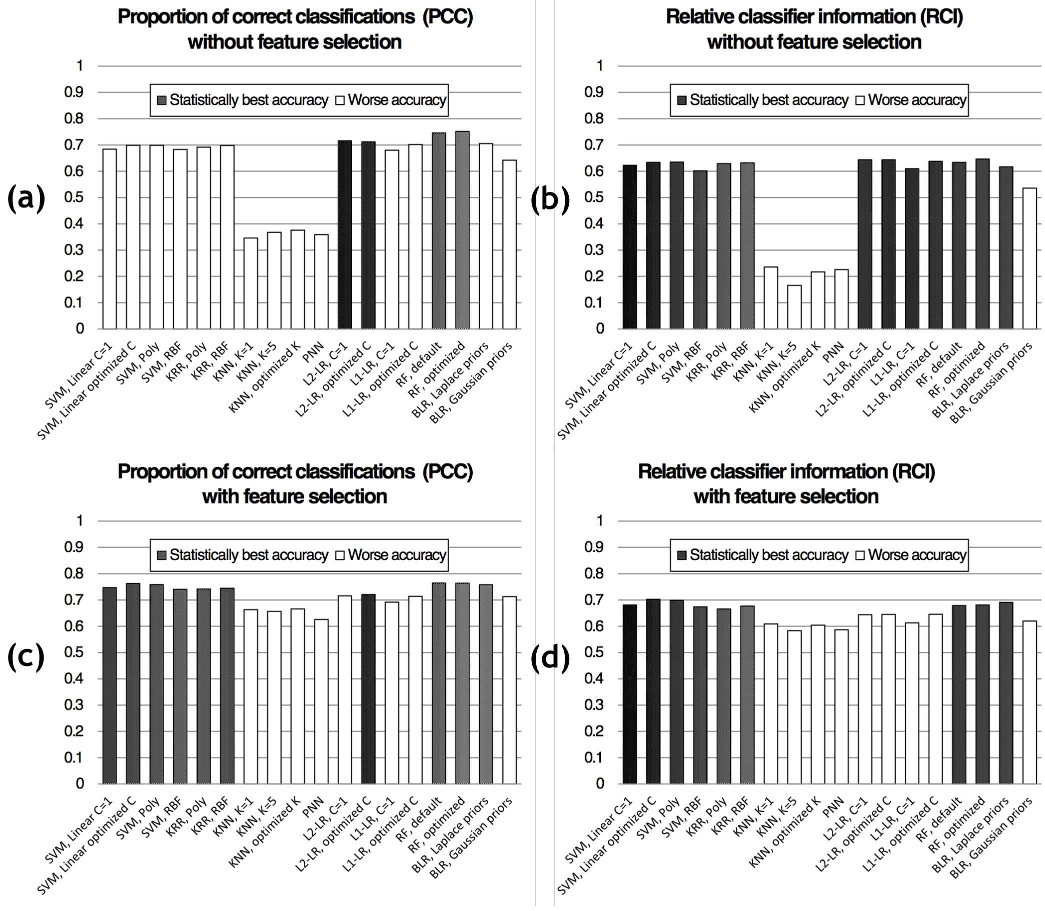
A comprehensive evaluation of multicategory classification methods for microbiomic data

Alexander Statnikov[1,2,§], Mikael Henaff[1], Varun Narendra[1], Kranti Konganti[5], Zhiguo Li[1], Liying Yang[2], Zhiheng Pei[2,3], Martin J. Blaser[2,4], Constantin F. Aliferis[1,3,6], Alexander V. Alekseyenko[1,2,§]

Problem: It is currently unknown which classifiers perform best among the many available alternatives for classification with microbiomic data linking abundances of microbial taxa to phenotypic and physiological states, which can inform development of new diagnostic, personalized medicine, and forensic modalities

Results: In this work, we performed a systematic comparison of 18 major classification methods, 5 feature selection methods, and 2 accuracy metrics using 8 datasets spanning 1,802 human samples and various classification tasks: body site and subject classification and diagnosis.

Conclusions: We found that random forests, support vector machines, kernel ridge regression, and Bayesian logistic regression with Laplace priors are the most effective machine learning techniques for performing accurate classification from these microbiomic data.



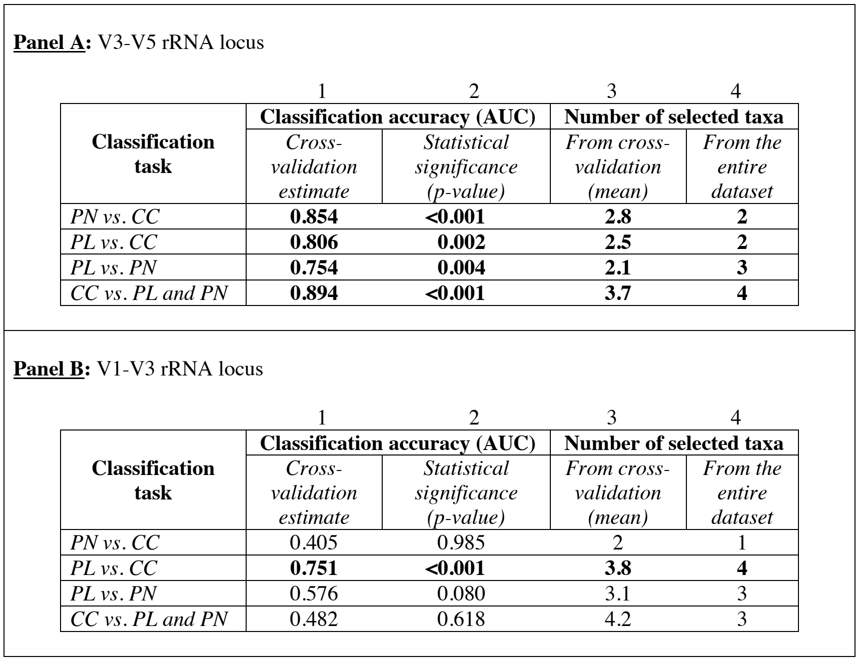
Comparison of State of the Art Methods for Microbiomic marker + Signature Discovery 2

Microbiomic Signatures of Psoriasis: Feasibility and Methodology Comparison

Alexander Statnikov[1,2,§], Alexander V. Alekseyenko[1,2], Zhiguo Li[1], Mikael Henaff[1], Guillermo I. Perez-Perez[2,3], Martin J. Blaser[2,3,5], Constantin F. Aliferis[1,4,6, §]

Problem: We sought to use bacterial community abundance data to develop multivariate molecular signatures of psoriasis for differentiation of cutaneous psoriatic lesions, clinically unaffected contralateral skin from psoriatic patients, and similar cutaneous loci in matched healthy control subjects. Using 16S rRNA high-throughput DNA sequencing, we assayed the cutaneous microbiome for 51 such triplet specimen including subjects of both genders, different age groups (18-81 years old) and ethnicities, and multiple body sites. We then assessed feasibility of multivariate molecular signatures to diagnose psoriasis

Results: it is possible to develop accurate molecular signatures for diagnosis of psoriasis from microbiomic data. The accuracy of molecular signatures depends on both DNA sequencing and downstream analysis protocols.

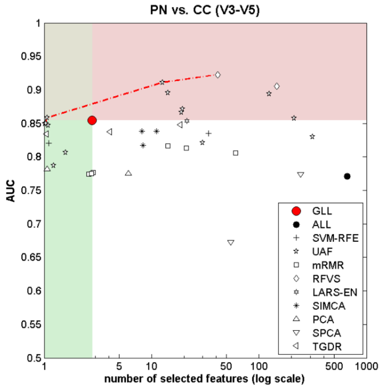


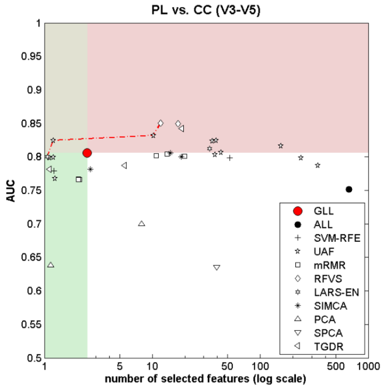
Panel A: V3-V5 rRNA locus

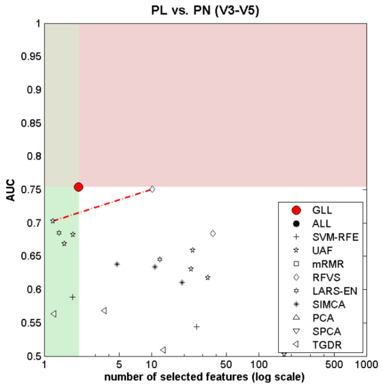
	1	2	3	4
Classification task	**Classification accuracy (AUC)**		**Number of selected taxa**	
	Cross-validation estimate	*Statistical significance (p-value)*	*From cross-validation (mean)*	*From the entire dataset*
PN vs. CC	**0.854**	**<0.001**	**2.8**	**2**
PL vs. CC	**0.806**	**0.002**	**2.5**	**2**
PL vs. PN	**0.754**	**0.004**	**2.1**	**3**
CC vs. PL and PN	**0.894**	**<0.001**	**3.7**	**4**

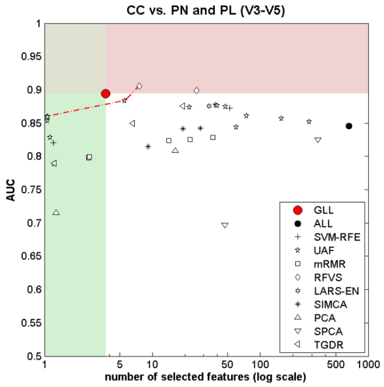
Panel B: V1-V3 rRNA locus

	1	2	3	4
Classification task	**Classification accuracy (AUC)**		**Number of selected taxa**	
	Cross-validation estimate	*Statistical significance (p-value)*	*From cross-validation (mean)*	*From the entire dataset*
PN vs. CC	0.405	0.985	2	1
PL vs. CC	**0.751**	**<0.001**	**3.8**	**4**
PL vs. PN	0.576	0.080	3.1	3
CC vs. PL and PN	0.482	0.618	4.2	3





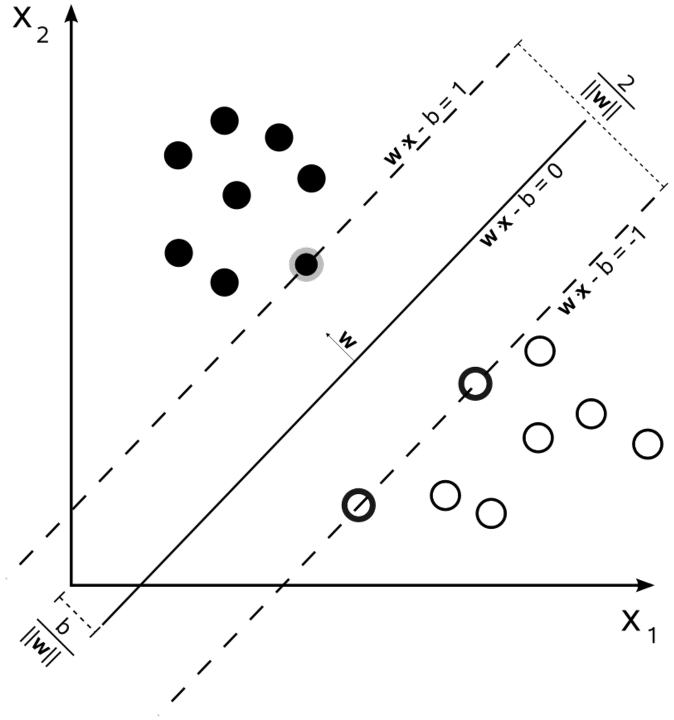




Classification techniques

- Support Vector Machines
- Random Forests

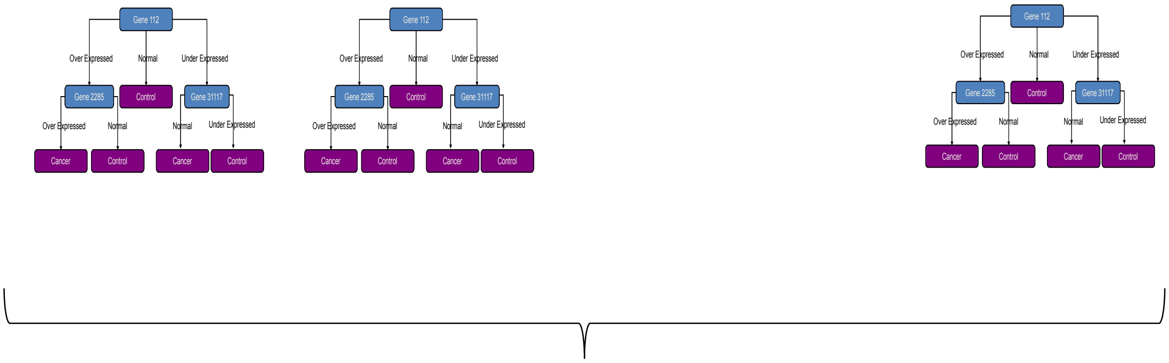
Support Vector Machines



X_2
$\mathbf{w}\cdot\mathbf{x}-b=1$
$\frac{2}{\|\mathbf{w}\|}$
$\mathbf{w}\cdot\mathbf{x}-b=0$
$\mathbf{w}\cdot\mathbf{x}-b=-1$
$\mathbf{w}$
$\frac{b}{\|\mathbf{w}\|}$
X_1

http://en.wikipedia.org/wiki/File:Svm_max_sep_hyperplane_with_margin.png

Random Forest



How well Random Forests perform in practice against state of the art methods such as SVM?

1. RFs perform well, almost on par with SVMs in terms of predictive accuracy.
2. RFs are slower than SVMs for typical HD molecular datasets.
3. RFs do not require to set up variable selection, model selection and error estimation separately because they embed those.
4. RFs often produce large, complicated, hard to explain models.