Midwest Big Data Summer School: Machine Learning II: Basic to Advanced Methods

Kris De Brabanter

kbrabant@iastate.edu

Iowa State University
Department of Statistics
Department of Computer Science

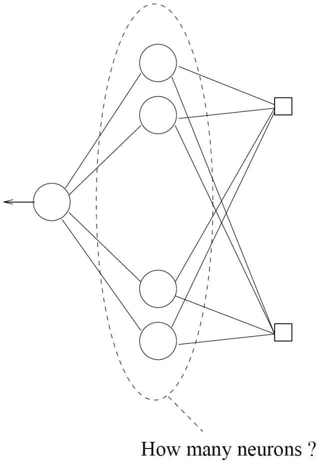
June 24, 2016

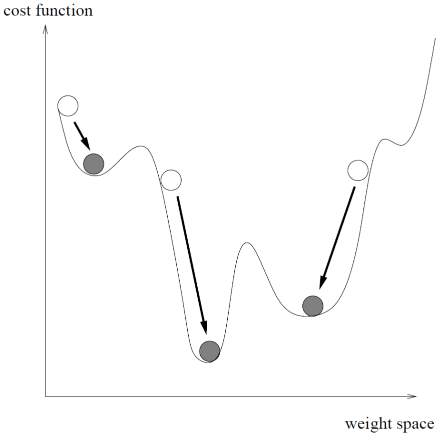
Outline

1. Support vector machines
    - Linearly separable case
    - Overlapping classes
    - Nonlinearly separable case

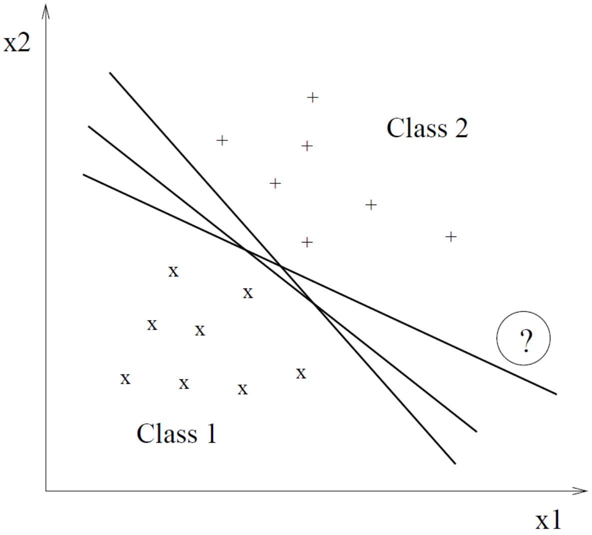
2. k-means clustering

Why support vector machines?



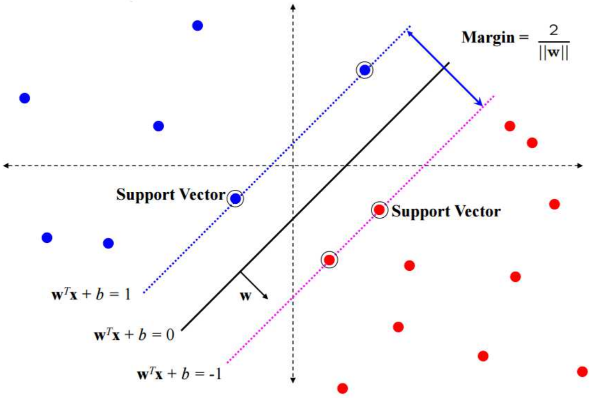


Why support vector machines?



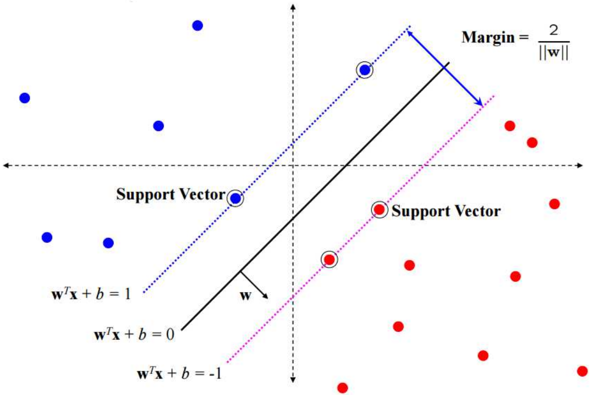
Main idea of support vector machines

Assume $\{X_i, Y_i\}_{i=1}^{n}$ with $X \in \mathbb{R}^p$ and $Y \in \{-1, 1\}$



Main idea of support vector machines

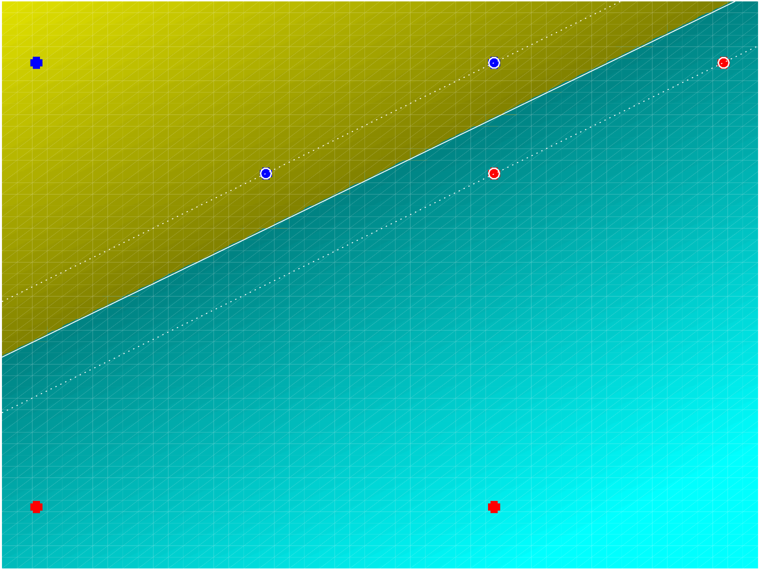
Assume $\{X_i, Y_i\}_{i=1}^n$ with $X \in \mathbb{R}^p$ and $Y \in \{-1, 1\}$



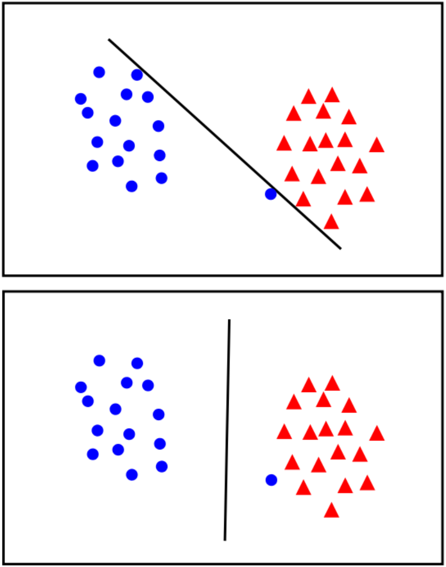
Need to solve the following constraint (convex) optimization problem

$$\min_w \frac{1}{2} w^T w \quad s.t. \quad Y_i(w^T X_i + b) \geq 1, i = 1, \ldots, n$$

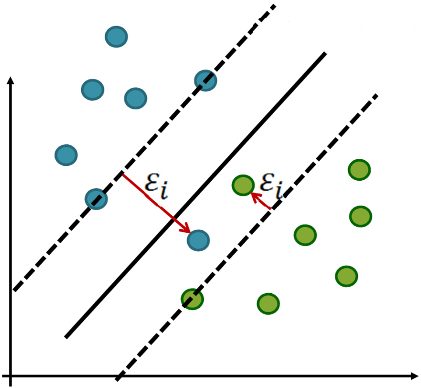
Example: classes are linearly separable



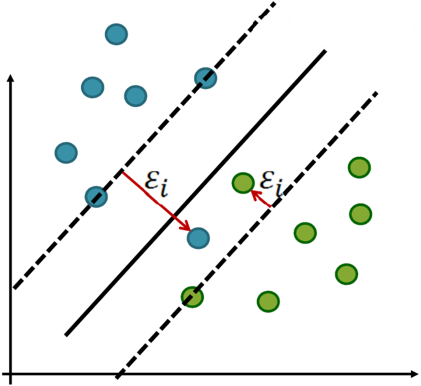
Overlapping classes



Slack variables



Slack variables

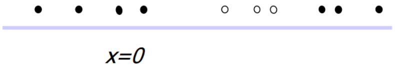


Need to solve the following constraint (convex) optimization problem

$$\min_w \frac{1}{2} w^T w + C \sum_{i=1}^{n} \varepsilon_i \quad s.t. \quad Y_i(w^T X_i + b) \geq 1 - \varepsilon_i, \ \varepsilon_i \geq 0, \ i = 1, \ldots, n$$

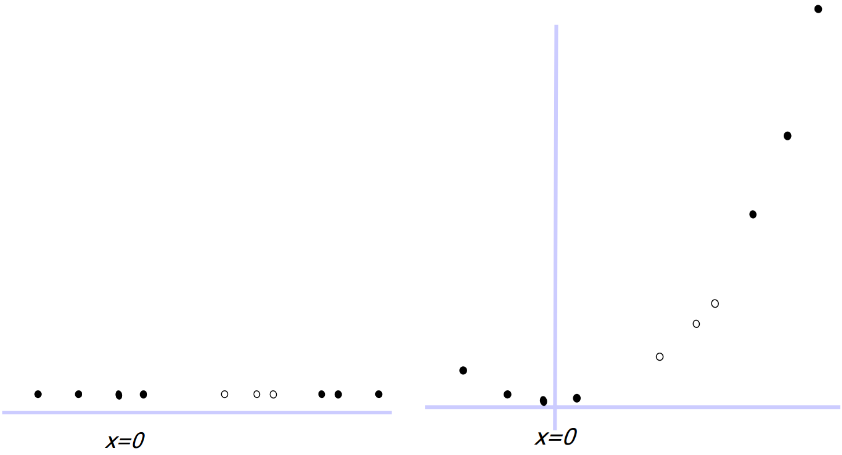
Nonlinearly separable

Solution: Map data into a higher dimensional (possibly infinite) space



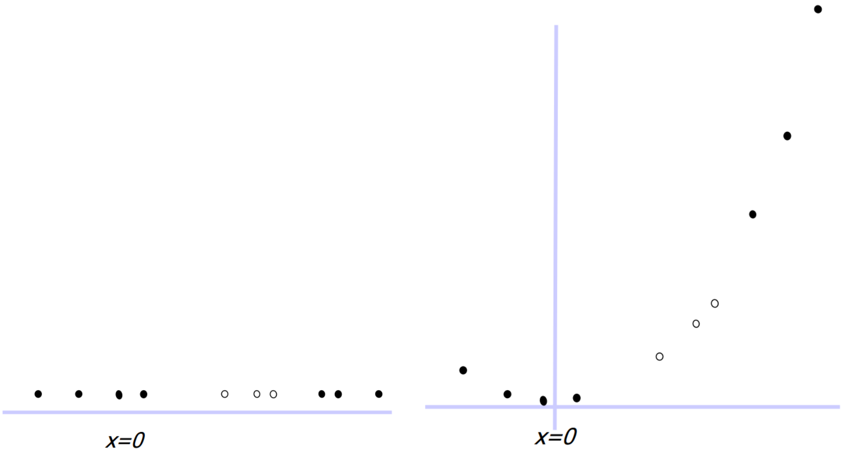
Nonlinearly separable

Solution: Map data into a higher dimensional (possibly infinite) space



Nonlinearly separable

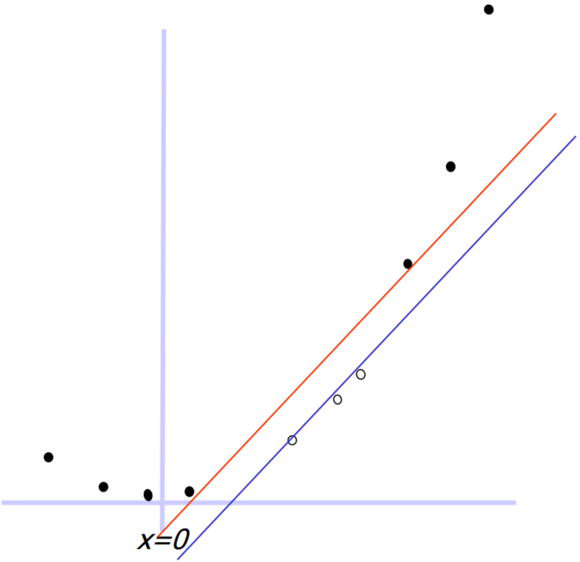
Solution: Map data into a higher dimensional (possibly infinite) space



Map data from 1D to 2D via the mapping $X \mapsto \varphi(X) = (X, X^2)$

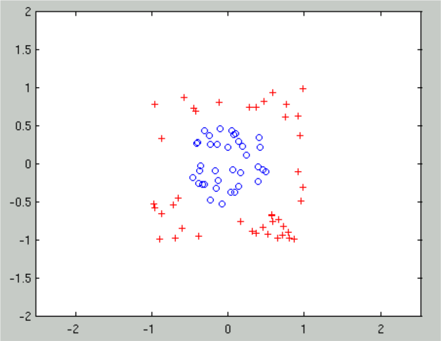
Mapping to high(er) dimensional space

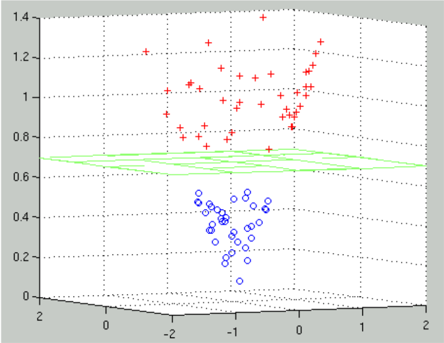
Data is linearly separable now. We can just use SVM in this new space



Mapping to high(er) dimensional space

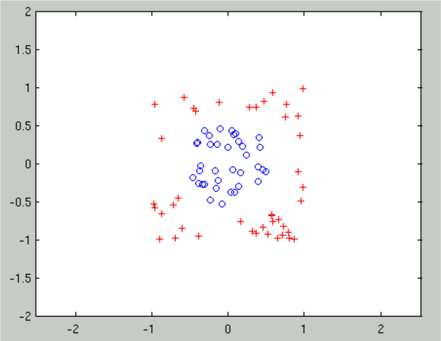
$$(X_1, X_2) \mapsto \varphi(X_1, X_2) = (X_1, X_2, \sqrt{X_1^2 + X_2^2})$$

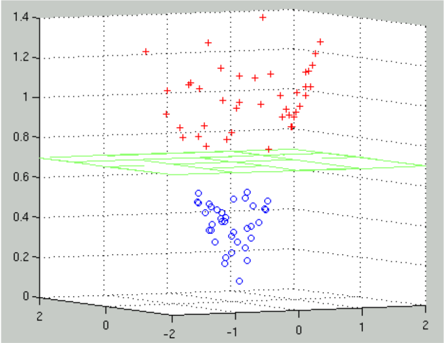




Mapping to high(er) dimensional space

$$(X_1, X_2) \mapsto \varphi(X_1, X_2) = (X_1, X_2, \sqrt{X_1^2 + X_2^2})$$





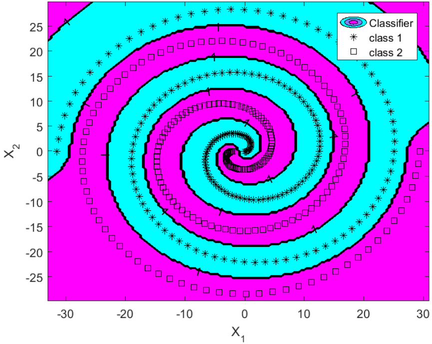
Need to solve the following constraint (convex) optimization problem

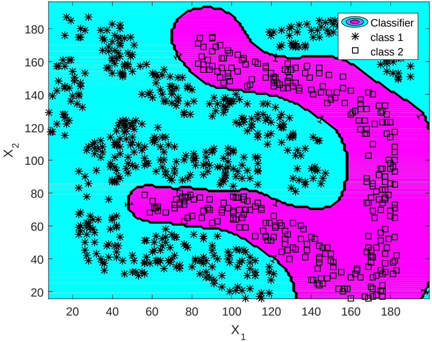
$$\min_w \frac{1}{2} w^T w + C \sum_{i=1}^{n} \varepsilon_i \quad s.t. \quad Y_i(w^T \varphi(X_i) + b) \geq 1 - \varepsilon_i, \ \varepsilon_i \geq 0, \ i = 1, \ldots, n$$

More on this mapping...

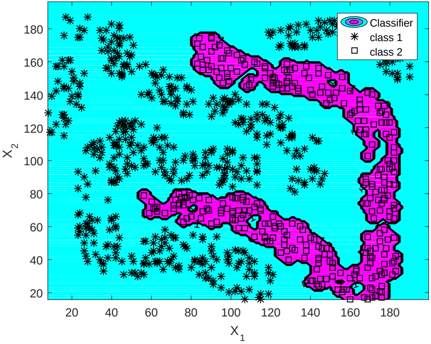
- Surprisingly you do not need to specify this mapping beforehand (Mercer, 1909)
- The inner product $\varphi(X_i)^T\varphi(X_j) = K(X_i, X_j)$ with K a positive definite kernel function
- Choices for K include linear, polynomial, Gaussian (with bandwidth h), etc.
- SVMs for pattern recognition usually have 1 or 2 tuning parameters (cross-validation)
- Related to Reproducing Kernel Hilbert Spaces

Two more nonlinear examples

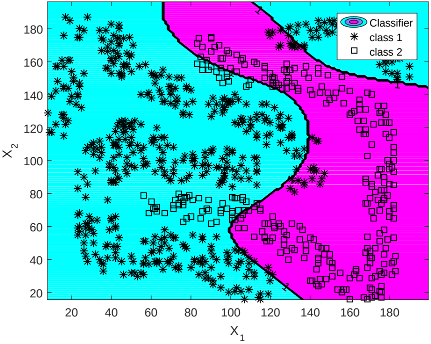




Importance of the tuning parameters



(i) h too small



(j) h too large

k-means clustering

Definition

Clustering refers to a very broad set of techniques for finding subgroups or clusters in a data set. It belongs into the category of unsupervised learning

k-means clustering

Definition

Clustering refers to a very broad set of techniques for finding subgroups or clusters in a data set. It belongs into the category of unsupervised learning

Given the total number of cluster K, let $C_1, \ldots, C_K$ denote sets containing indices of the observations in each cluster. These sets satisfy

1. $C_1 \cup \cdots \cup C_K$={1,. . . ,K}
2. $C_k \cap C_{k'} = \emptyset$ for all $k \neq k'$

k-means clustering

Definition

Clustering refers to a very broad set of techniques for finding subgroups or clusters in a data set. It belongs into the category of unsupervised learning

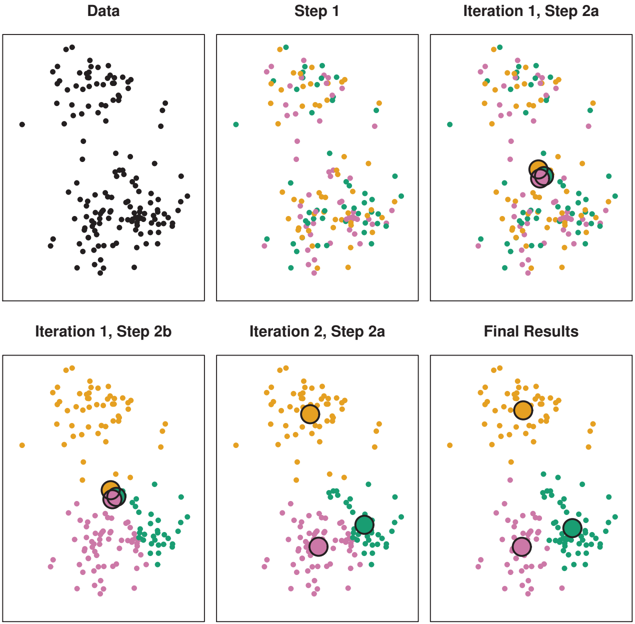
Given the total number of cluster K, let $C_1, \ldots, C_K$ denote sets containing indices of the observations in each cluster. These sets satisfy

1. $C_1 \cup \cdots \cup C_K = \{1, \ldots, K\}$
2. $C_k \cap C_{k'} = \emptyset$ for all $k \neq k'$

The idea behind k-means clustering is that "good" clustering is one for which the within-cluster variation is as small as possible

This turns out to be computationally infeasible since there are almost K^n ways to partition n observations into K clusters. Solution: Iterative algorithm that finds a local optimum!

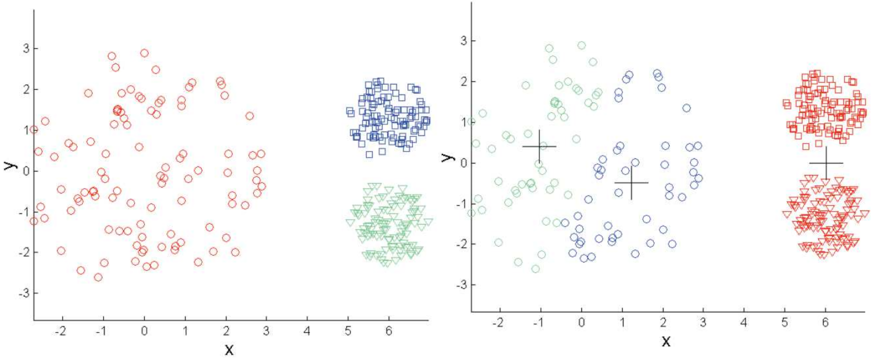
The k-means algorithm in action



Limitations of k-means

- k-means has problems when clusters are of different
    - sizes
    - densities
    - non-globular shapes
- problem with outliers
- empty clusters

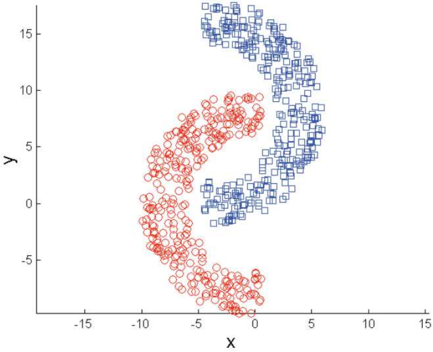
Limitations of k-means: Different densities



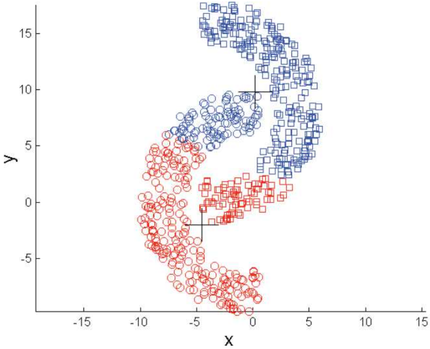
(a) original data

(b) k-means clustering with $k = 3$

Limitations of k-means: Non-globular shapes



(c) original data



(d) k-means clustering with $k = 2$

Extra info

How to determine k?

Extra info

How to determine k?

- Silhouettes (Rousseeuw, 1986)
- Gap Statistic
- See Milligan & Cooper (1985) for a comprehensive simulation comparison of 30 different procedures

Extra info

How to determine k?

- Silhouettes (Rousseeuw, 1986)
- Gap Statistic
- See Milligan & Cooper (1985) for a comprehensive simulation comparison of 30 different procedures

Other clustering methods

- Hierarchical clustering (does not require a priori knowledge of the number of clusters)
- Spectral clustering
- Ward's method

References

- Bühlmann, P & van de Geer S. (2011), Statistics for High-Dimensional Data: Methods, Theory and Applications, Springer
- Devroye L., Györfi L. & Lugosi G. (1996). A Probabilistic Theory of Pattern Recognition, Springer
- Gareth J., Witten D., Hastie T. & Tibshirani R. (2013). An Introduction to Statistical Learning with Applications in R, Springer
- Hastie T., Tibshirani R. & Friedman J. (2009), The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd Ed.), Springer
- Hoerl A.E. & Kennard R. (1974), Ridge regression: biased estimation for nonorthogonal problems, *Technometrics*, 12: 55-67
- Milligan G.W. & Cooper M.C. (1985). An examination of procedures for determining the number of clusters in a data set, Psychometrika 50:159-179
- Rousseeuw P. J. (1987). Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. Computational and Applied Mathematics 20: 5365
- Silverman B. W., Density Estimation for Statistics and Data Analysis, Chapman & Hall, 1986
- R. Tibshirani (1996), Regression shrinkage and selection via the lasso, *J. Royal. Statist. Soc B.*, 58(1): 267-288
- R. Tibshirani, G. Walther & T. Hastie (2001), Estimating the number of clusters in a data set via the gap statistic, J.R.Statist. Soc. B. 63(2): 411-423