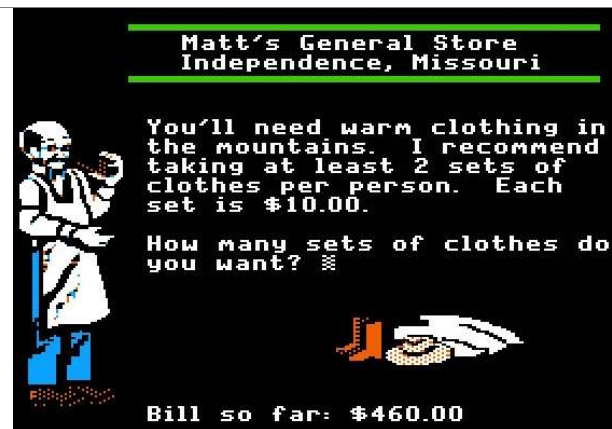

Classification and Regression Trees



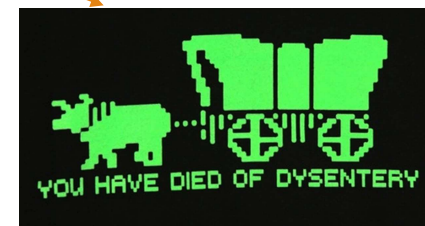
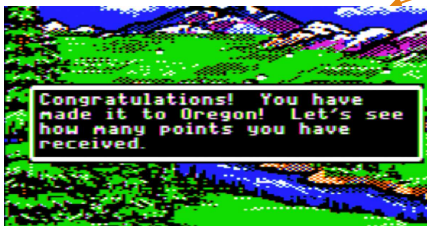
Intro to Tree-Based Methods: Will you die of dysentery?



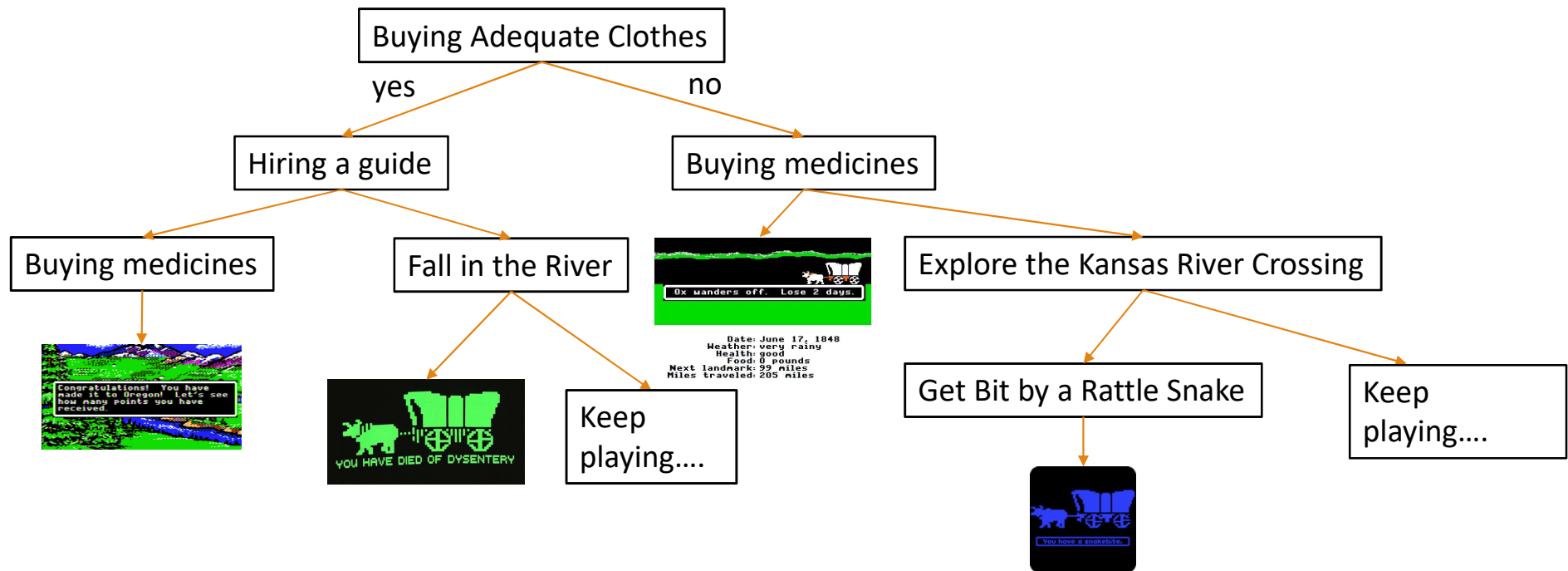
Yes

Buy 2 sets of clothes

No

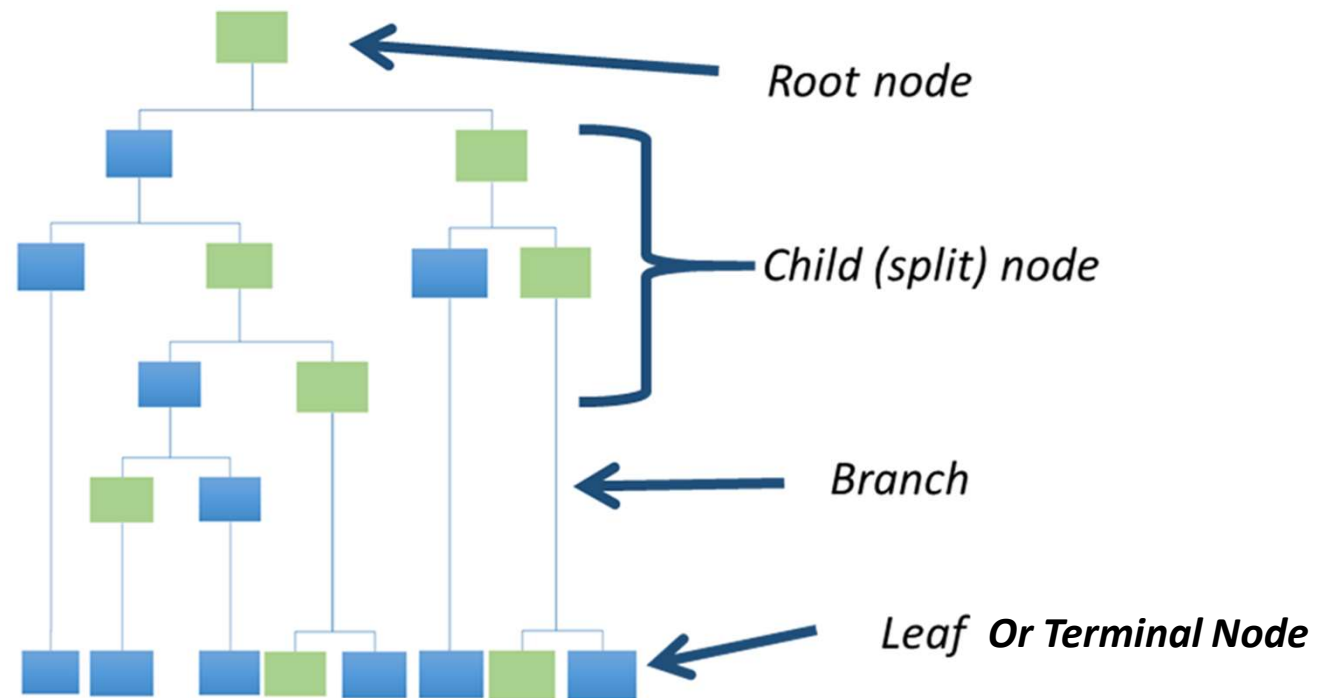


If we were nerdier....



Structure of a Tree

Trees generated through recursive partitioning



Key Terms

“Greedy” Algorithm: makes the optimal choice at each step

- Makes “best” first split without consideration of subsequent splits

Surrogate split: split using another feature that most closely resembles the consequences of the original split

- Often how tree-based methods handle missing data

Node purity: homogeneity of a node in relation to the labels of the observations contained

- Goal is often to maximize node purity to obtain optimal classification or prediction

Measures of purity


- Gini coefficient, entropy, variance, mean square error.... and others

Decision, Classification, Regression Trees: What's in a name?

- **Decision trees:** general name that can describe diverse applications of tree-structures
 - Trees for prediction, common in data analytics
 - Trees for decision analysis, common in business and engineering applications
- **Classification and Regression Trees (CaRT):** term introduced by Breiman et al in 1984
 - Classification: outcome that is the target of prediction is binary/categorical
 - Regression: outcome that is the target of prediction is continuous
 - Structure of trees similar, but different criteria for splitting and evaluation

Examples in Epidemiology

Using the PDD Behavior Inventory as a Level 2 Screener: A Classification and Regression Trees Analysis

Ira L. Cohen¹  · Xudong Liu² · Melissa Hudson² · Jennifer Gillis³ ·
Rachel N. S. Cavalari³ · Raymond G. Romanczyk³ · Bernard Z. Karmel⁴ ·
Judith M. Gardner⁴

**Applications often extend beyond traditional
prediction problems.**

Risk stratification for 25-year cardiovascular disease incidence in type 1 diabetes: Tree-structured survival analysis of the Pittsburgh Epidemiology of Diabetes Complications study

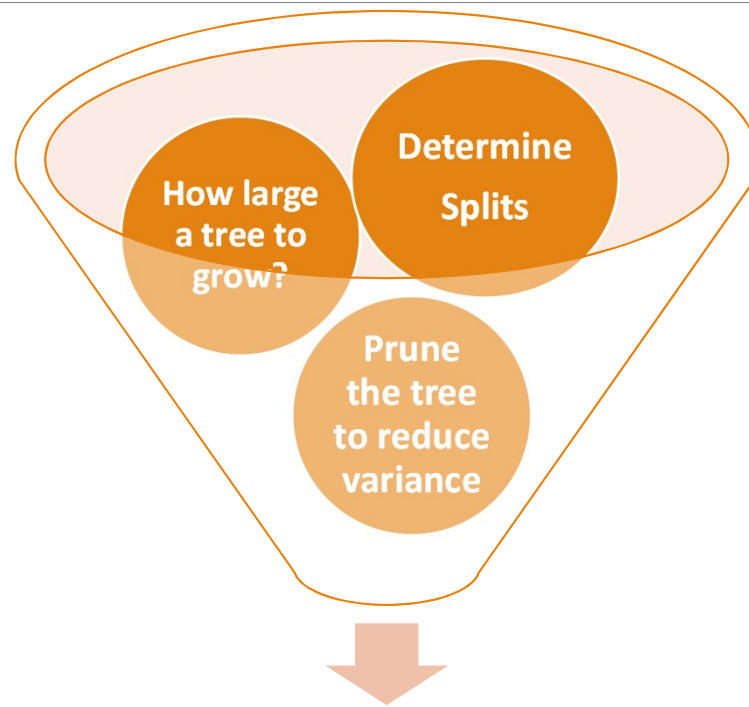
Rachel G Miller¹, Stewart J Anderson², Tina Costacou¹,
Akira Sekikawa¹ and Trevor J Orchard¹

Original article

Potential selection bias associated with using geocoded birth records for epidemiologic research

Sandie Ha PhD, MPH ^a, Hui Hu BS ^a, Liang Mao PhD ^b, Dikea Roussos-Ross MD ^c, Jeffrey Roth PhD ^d, Xiaohui Xu
PhD ^e  

Growing a Tree: Analytic Considerations



Constructing a tree-based model

Split Criteria

General Goal: Maximize node purity

Step 1: Divide the Feature Space

For each feature, we examine each potential split that partitions the full data into two subsets.

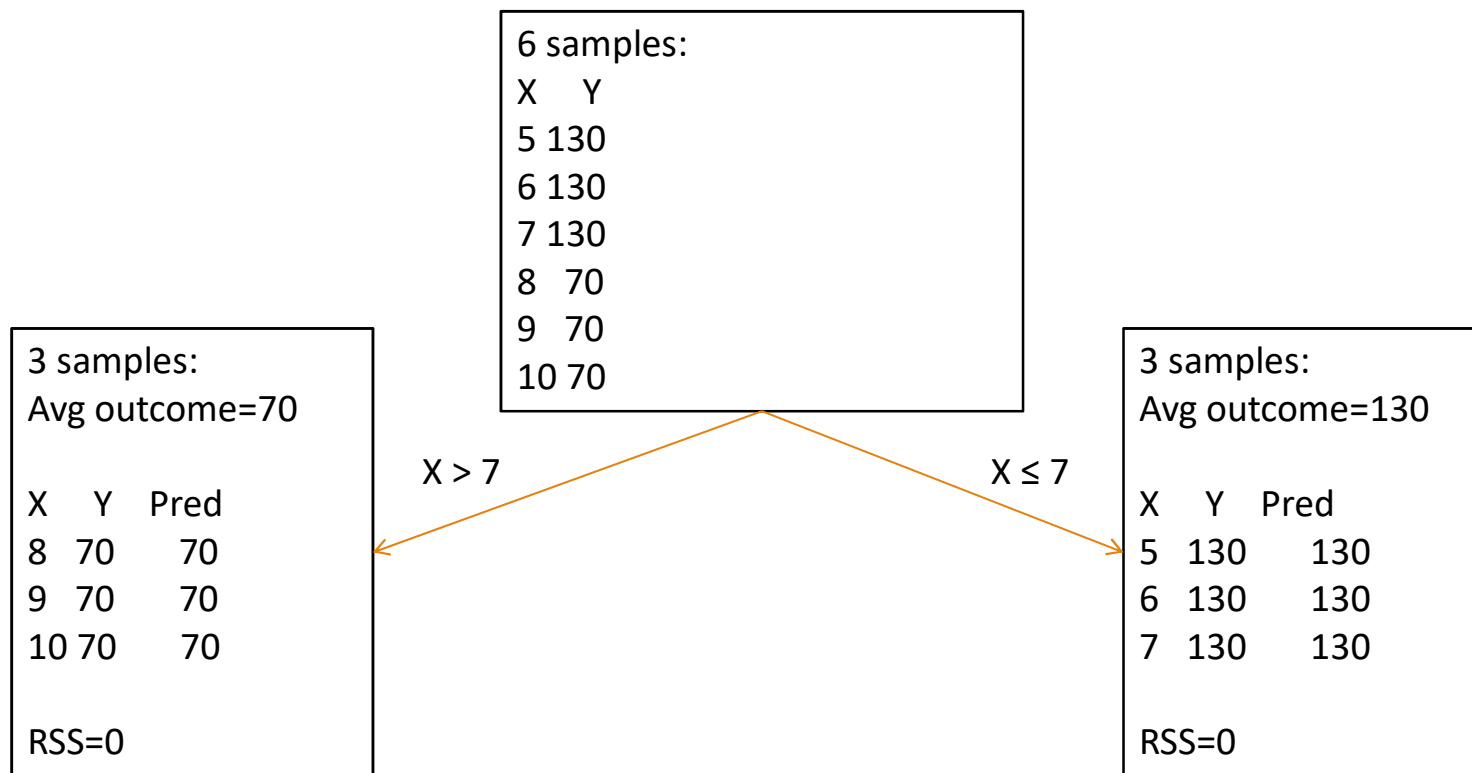
Step 2: Assign average outcome value/modal classification as prediction to all observations within subset

Feature-split combination that leads to the greatest purity is selected

Regression Trees

Purity: minimize the residual sum of squares

Sample splits to illustrate purity



Split Criteria

General Goal: Maximize node purity

Step 1: Divide the Feature Space

For each feature, we examine each potential split that partitions the full data into two subsets.

Step 2: Assign average outcome value/modal classification as prediction to all observations within subset

Feature-split combination that leads to the greatest purity is selected

Regression Trees

Purity: minimize the residual sum of squares

Classification Trees

Purity: Classification error rate, Gini, Entropy.

Error Rate: 1 - proportion of training observations in the subset (m) that have the predicted class (k) p_{mk}

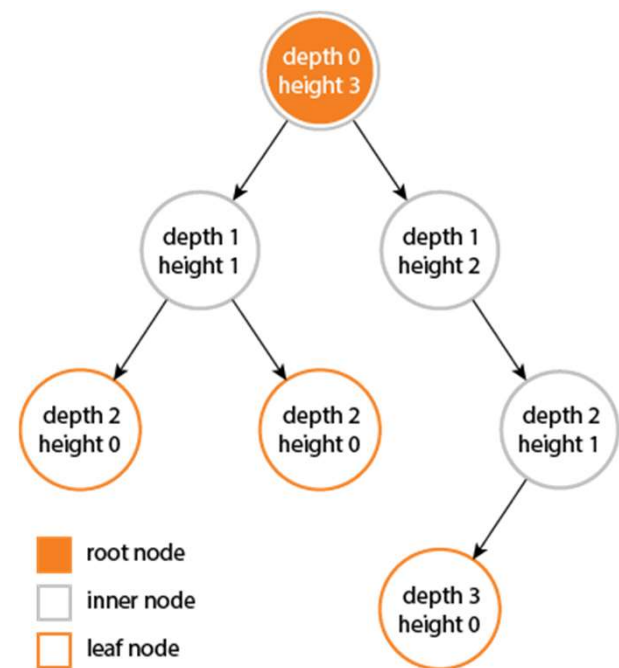
Gini impurity: $G = \sum_{k=1}^K p_{mk}(1 - p_{mk})$ measure of error across the different predicted classes

REPEAT THIS PROCESS RECURSIVELY UNTIL STOPPING CRITERION HAS BEEN REACHED

Criteria for stopping tree growth

Typically, grow a large tree and then prune back

- Maximum tree depth
- Minimum number of observations to consider potential splits
- Minimum number of observations in terminal node
- Lack of improvement in node purity (using *a priori threshold*)



Pruning

Process of reducing the size of the tree to avoid overfitting

Seeks balance between its fit to the data (bias) and reducing complexity (variance)

Cp: complexity parameter (typically denoted as α)

- Can think of it as a penalty for increasing the size/complexity of the tree

Equation for subtree:

$$\sum_{m=1}^{|T|} \sum_{i: x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T|$$

$|T|$: number of terminal nodes; R_m is subset of feature space

As cp (or α) increases from 0, branches get pruned from the tree in a nested and predictable fashion. Often selected via cross-validation methods to minimize error (\sim within 1 SD).

Variable Importance

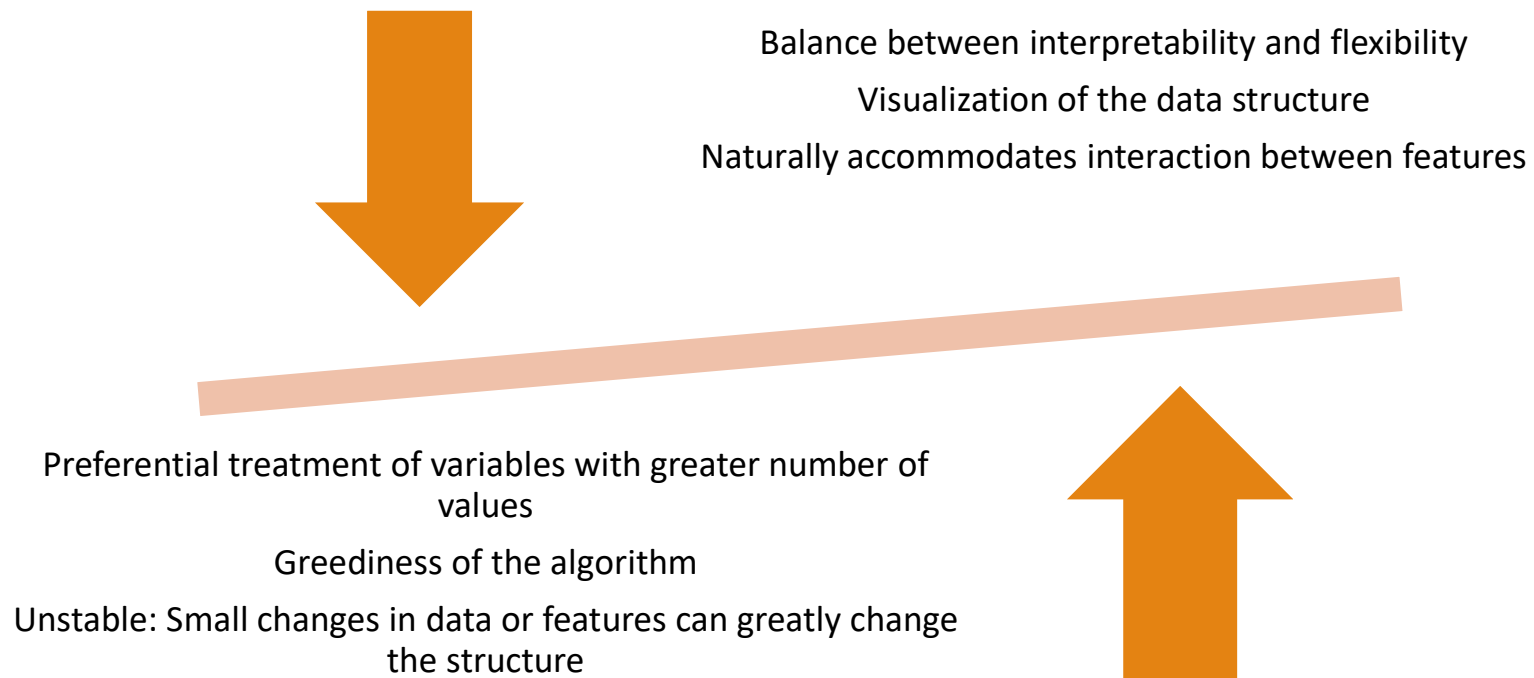
Interpretability is a strength of tree-based methods

Variable importance: metric to reflect a variable's importance to the prediction

- Often interpreted as a ranking rather as a quantitative measure
- Calculation depends upon the algorithm used, so check documentation

From Rpart documentation: “A variable may appear in the tree many times, either as a primary or a surrogate variable. An overall measure of variable importance is the sum of the goodness of split measures for each split for which it was the primary variable, plus goodness * (adjusted agreement) for all splits in which it was a surrogate. In the printout these are scaled to sum to 100 and the rounded values are shown, omitting any variable whose proportion is less than 1%. Imagine two variables which were essentially duplicates of each other; if we did not count surrogates they would split the importance with neither showing up as strongly as it should.”

Strengths and Limitations

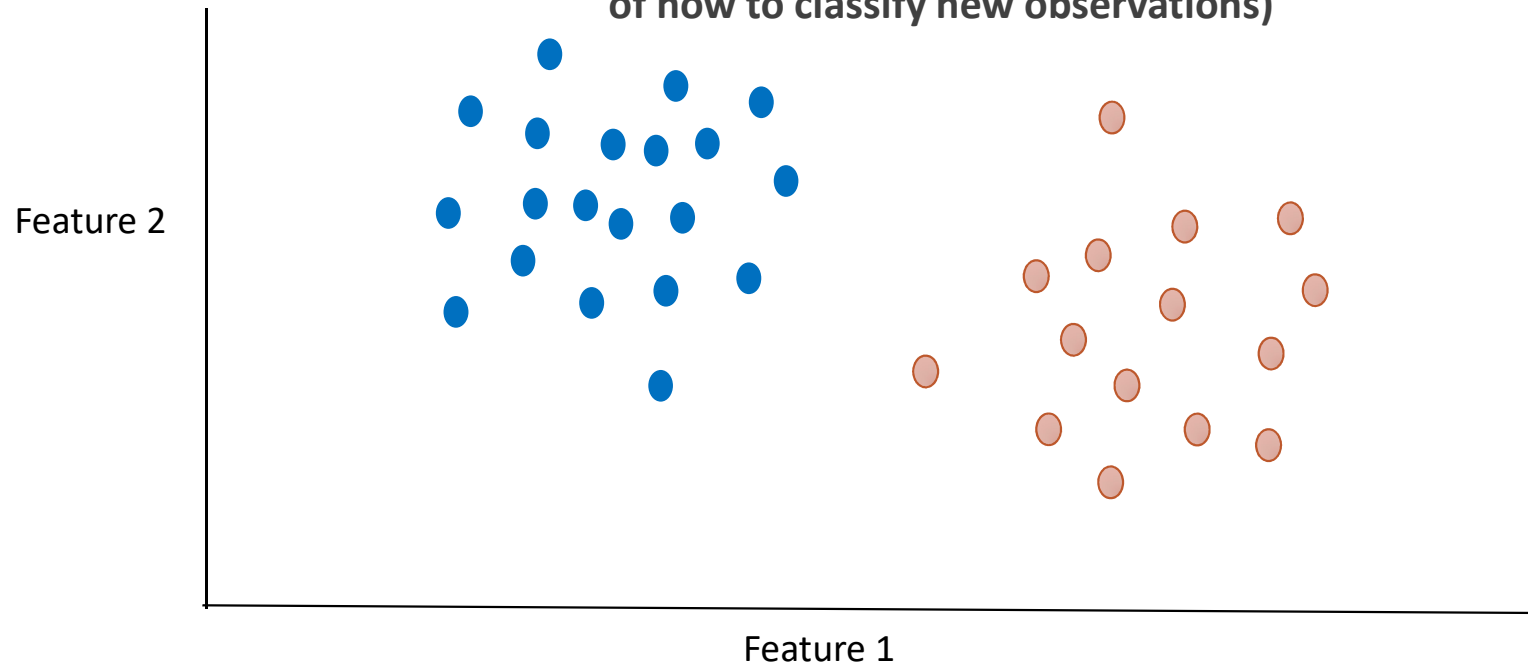


Support Vector Machines

A solid orange horizontal bar spanning the width of the slide, located at the bottom.

Example of Linear Separation

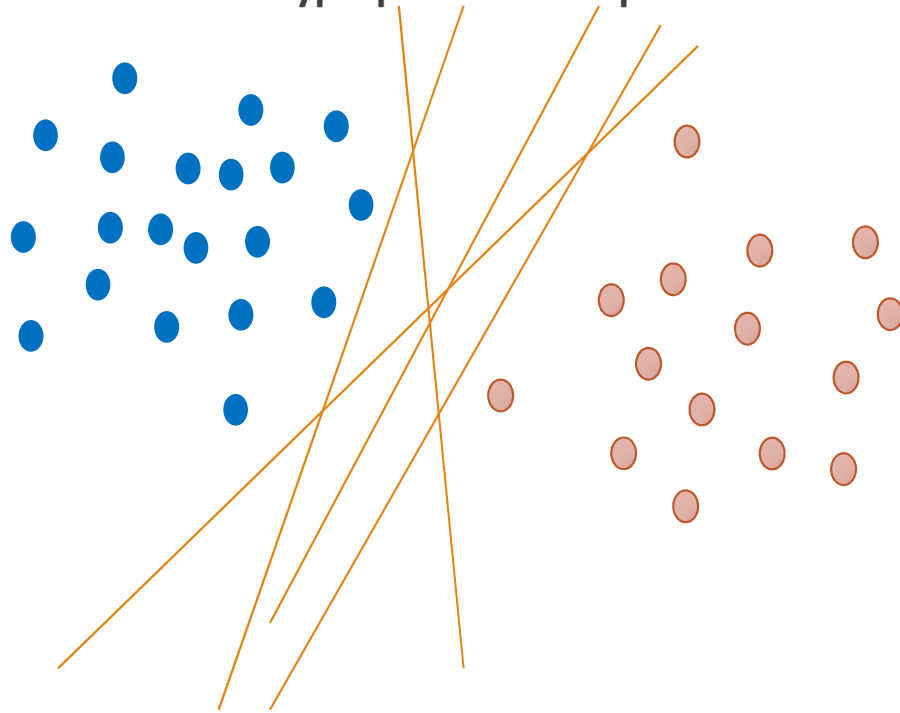
Goal of classification seeks a separation of data into distinct classes (and provides instruction of how to classify new observations)



Separate using a hyperplane (line in 2-D)

Hyperplane: subspace whose dimension is one less than its surrounding space

Infinite number of hyperplanes can separate these two classes



Choice of Hyperplane dictates Prediction

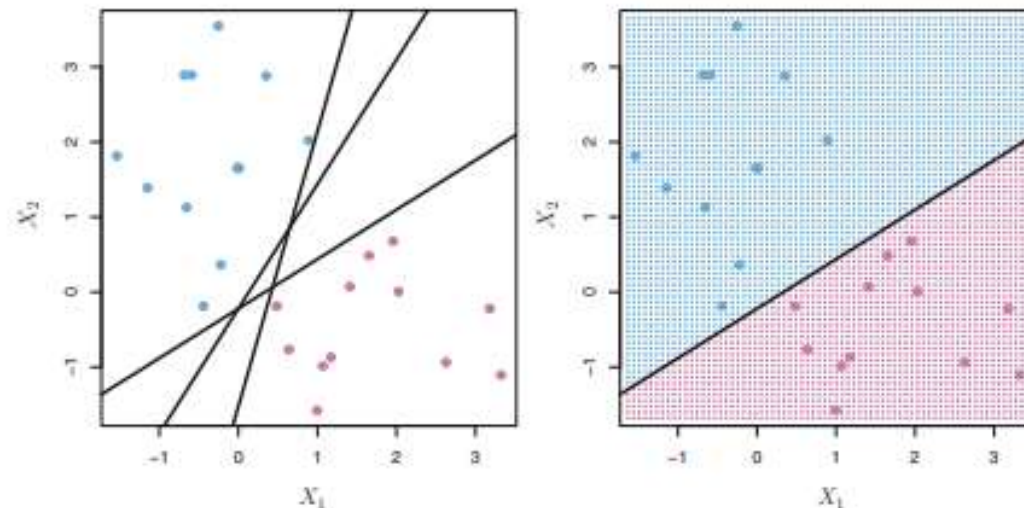


FIGURE 9.2. Left: There are two classes of observations, shown in blue and in purple, each of which has measurements on two variables. Three separating hyperplanes, out of many possible, are shown in black. Right: A separating hyperplane is shown in black. The blue and purple grid indicates the decision rule made by a classifier based on this separating hyperplane: a test observation that falls in the blue portion of the grid will be assigned to the blue class, and a test observation that falls into the purple portion of the grid will be assigned to the purple class.

Choice of Hyperplane dictates Prediction

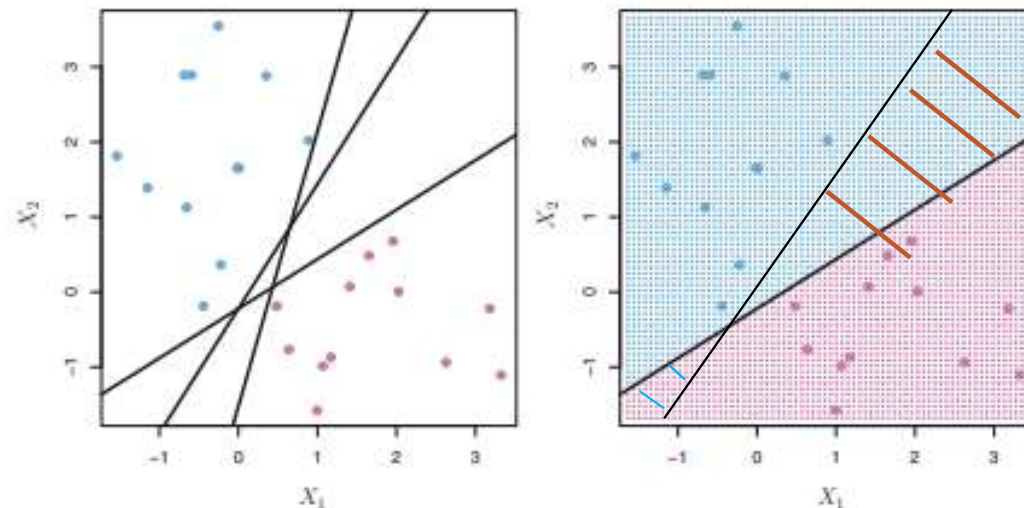


FIGURE 9.2. Left: There are two classes of observations, shown in blue and in purple, each of which has measurements on two variables. Three separating hyperplanes, out of many possible, are shown in black. Right: A separating hyperplane is shown in black. The blue and purple grid indicates the decision rule made by a classifier based on this separating hyperplane: a test observation that falls in the blue portion of the grid will be assigned to the blue class, and a test observation that falls into the purple portion of the grid will be assigned to the purple class.

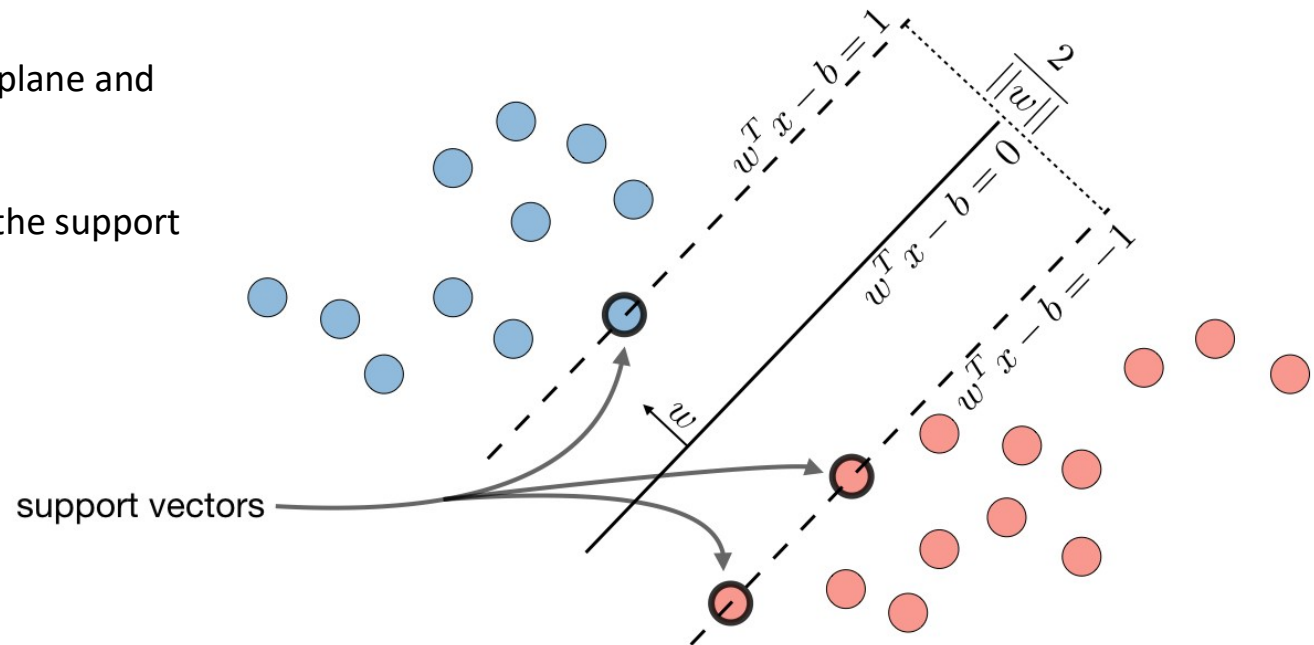
Support Vectors

Support vectors

Observations that lie closest to the hyperplane and are often the most difficult to classify

Goal is to maximize the margin between the support vectors

Solved via optimization techniques

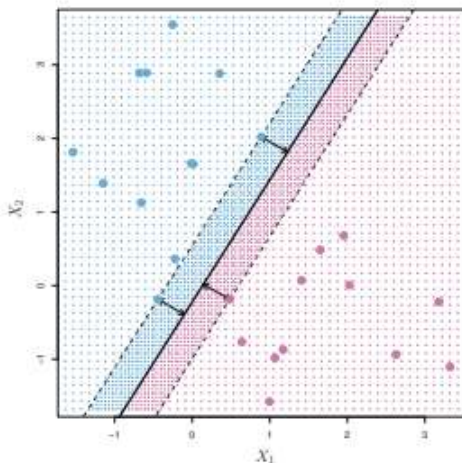


Confusion of Terminology: Not all SVC

Maximal Margin Classifier: when a hyperplane exists to perfectly separate the classes

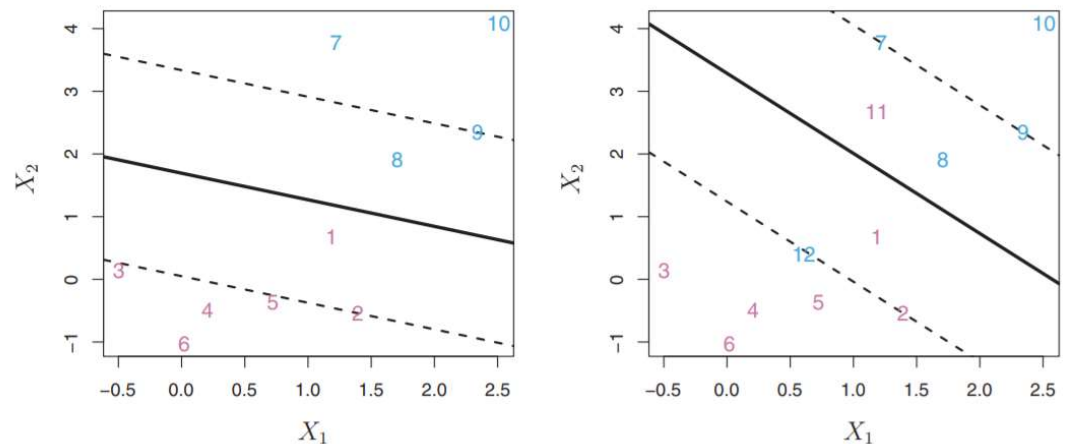
Support Vector Classifier: extension of maximal margin classifier when hyperplane does not perfectly separate the classes (aka soft margin classifier)

Maximal margin classifier



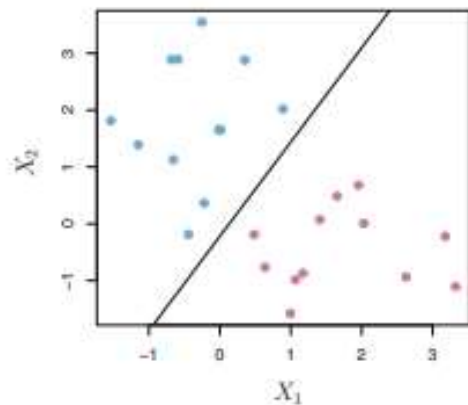
Source: ISLR Fig 9.3

Support vector classifier

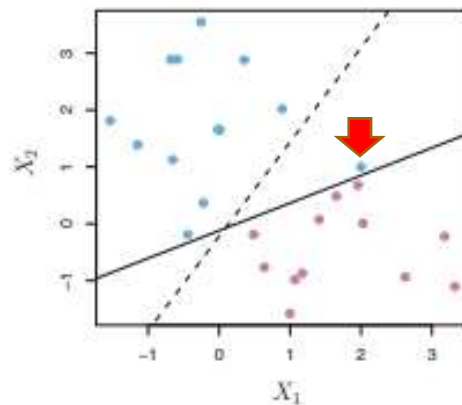


Source: ISLR Fig 9.6

Why violate the margins?



Initial maximal margin



Adding a new point changes
everything

**Using support vector classifier,
instead of maximal margin classifier:**

- Greater robustness to individual observations (greater stability)
- Better classification of most training observations

REMEMBER FOR PROGRAMMING

C: tuning parameter that controls how much margin can be violated (i.e. how much misclassification)
C=0, no violations and equivalent to Maximal Margin Classifier

Reminder of Key Terminology

Support Vector Machine sometimes used as umbrella term but technically:

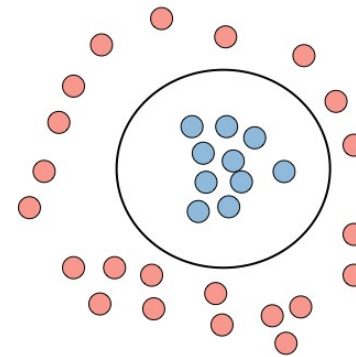
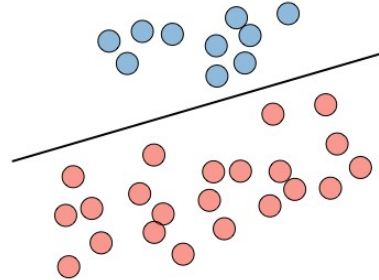
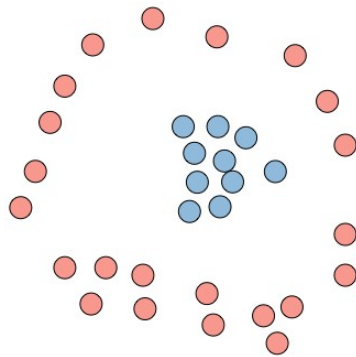
Maximal Margin Classifier: when a hyperplane exists to perfectly separate the classes

Support Vector Classifier: extension of maximal margin classifier when hyperplane does not perfectly separate the classes (aka soft margin classifier)

Support Vector Machine: extension of the support vector classifier when the classes are not linearly separable in the current feature space

Accommodation of Non-Linearity

Imagine I'm rotating the points



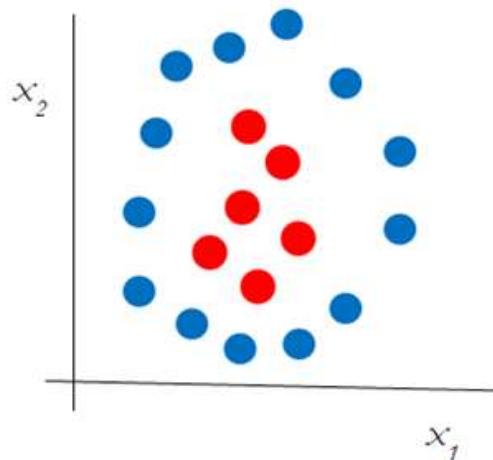
Non-linear separability \longrightarrow Use of a kernel mapping ϕ \longrightarrow Decision boundary in the original space

\uparrow

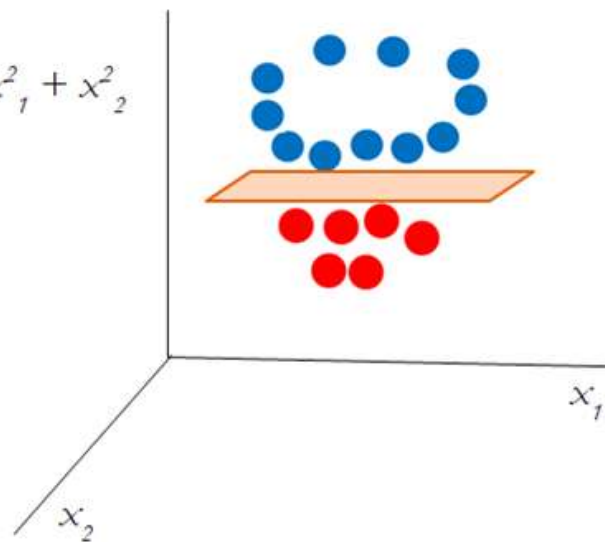
Projecting onto a space where the problem is linearly separable

Another visual of kernel projection

2-Dimensional Linearly Inseparable Classes



2-Dimensional Linearly Inseparable Classes with Polynomial kernel with Degree 2



Applications in Epidemiology

RESEARCH ARTICLE

Morbidity Rate Prediction of Dengue Hemorrhagic Fever (DHF) Using the Support Vector Machine and the *Aedes aegypti* Infection Rate in Similar Climates and Geographical Areas

Kraisak Kesorn , Phatsavee Ongruk, Jakkrawarn Chomposri, Atchara Phumee, Usavadee Thavara, Apiwat Tawatsin, Padet Siriyasatien

[J Health Care Poor Underserved. 2013 Feb; 24\(1.0\): 153–171.](#)

doi: [10.1353/hpu.2013.0046](#)

Analysis of an Environmental Exposure Health Questionnaire in a Metropolitan Minority Population Utilizing Logistic Regression and Support Vector Machines

[Chau-Kuang Chen](#), EdD, [Michelle Bruce](#), MD, MSPH, [Lauren Tyler](#), BS, [Claudine Brown](#), MSPH, [Angelica Garrett](#), MD, [Susan Goggins](#), MD, [Brandy Lewis-Polite](#), MD, [Mirabel L Weriwoh](#), MD, MSPH, [Paul D. Juarez](#), PhD, [Darryl B. Hood](#), PhD, and [Tyler Skelton](#), MS

BIOINFORMATICS

Vol. 27 ISMB 2011, pages i342–i348
doi:10.1093/bioinformatics/btr204

ccSVM: correcting Support Vector Machines for confounding factors in biological data classification

Limin Li^{1,2,*}, Barbara Rakitsch¹ and Karsten Borgwardt^{1,*}

¹Machine Learning and Computational Biology Research Group, Max Planck Institutes Tübingen, Tübingen, Germany and ²Department of Mathematics, Xi'an Jiaotong University, Xi'an 710049, China

Relationship between SVC and Logistic Regression

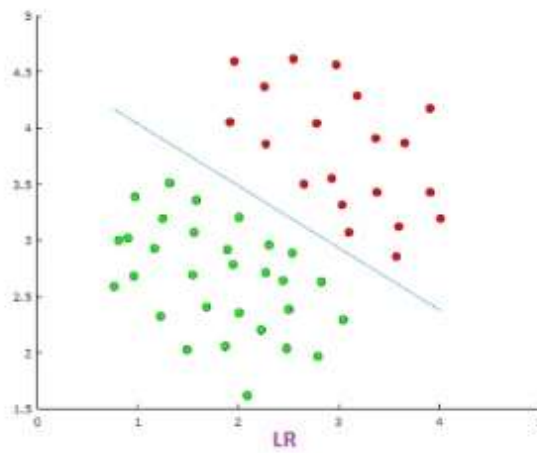
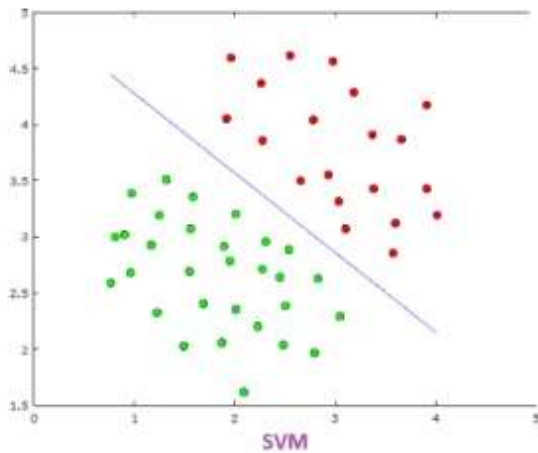
Only slight differences in the loss functions

$$\min_w \lambda \|w\|^2 + \sum_i \max\{0, 1 - y_i w^T x_i\}$$

Loss function for Support Vector Classifier

$$\min_w \lambda \|w\|^2 + \sum_i \log(1 + \exp(1 - y_i w^T x_i))$$

Loss function for Logistic Regression (regularized)



SVM: maximizes distance between classes

LR: maximizes posterior class probability (can think of it as focused on distance from one side)

Summary

- Can be used for classification or regression
- Typical implementation is support vector machine with linear kernel -> support vector classifier
 - Similar(theoretically slightly better) results than logistic regression
- Limited interpretability
 - Can utilize permutation to learn about importance of single variable