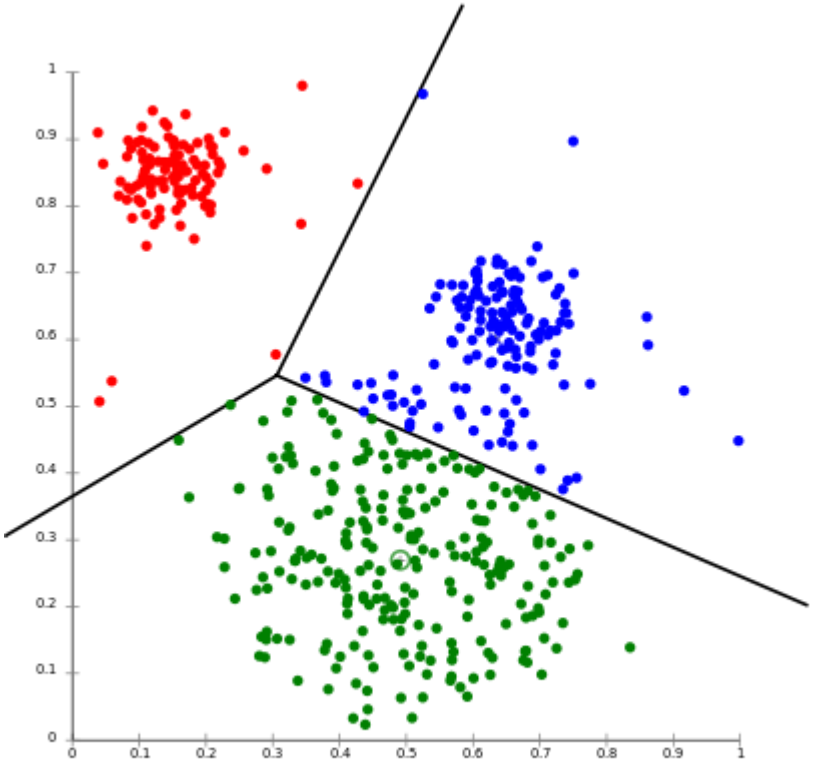
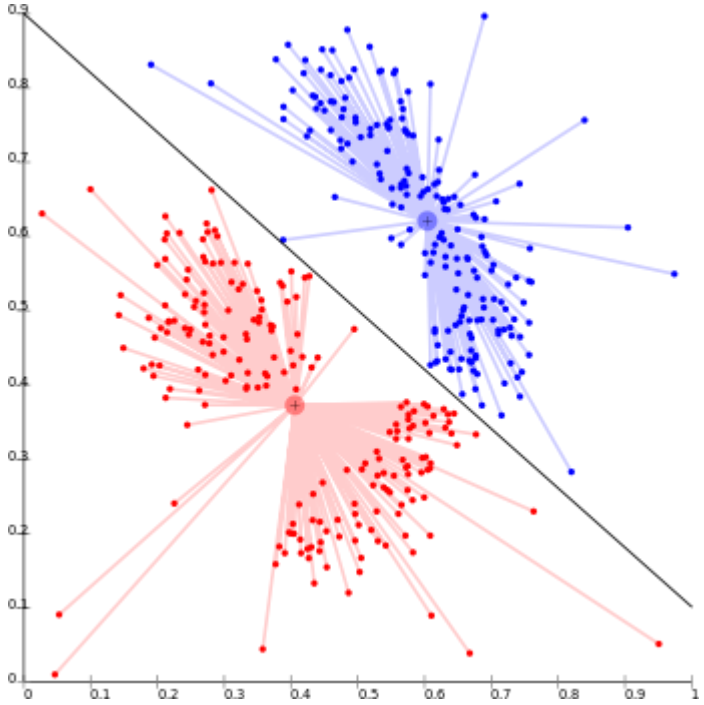
Biology 644

Old Title: Bioinformatics for Molecular Biologists

Potential New Title: Integrated Bioinformatics Using R for Both Wet and Dry Scientists

Remember how Centroid and Distribution-based Clustering can go wrong?





Support Vector Machines to the Rescue

- A Support Vector Machine (SVM) performs classification by constructing an N-dimensional hyperplane that optimally separates the data into two categories.

- SVM models are closely related to neural networks, but more flexible and understandable.
  - an SVM model using a sigmoid kernel is equivalent to a two-layer, neural network
  - In feature space, SVMs solve a quadratic programming problem with linear constraints, instead of solving a non-convex, unconstrained minimization problem

- In SVM literature, a predictor variable is called an attribute, and a transformed attribute that is used to define the hyperplane is called a feature.

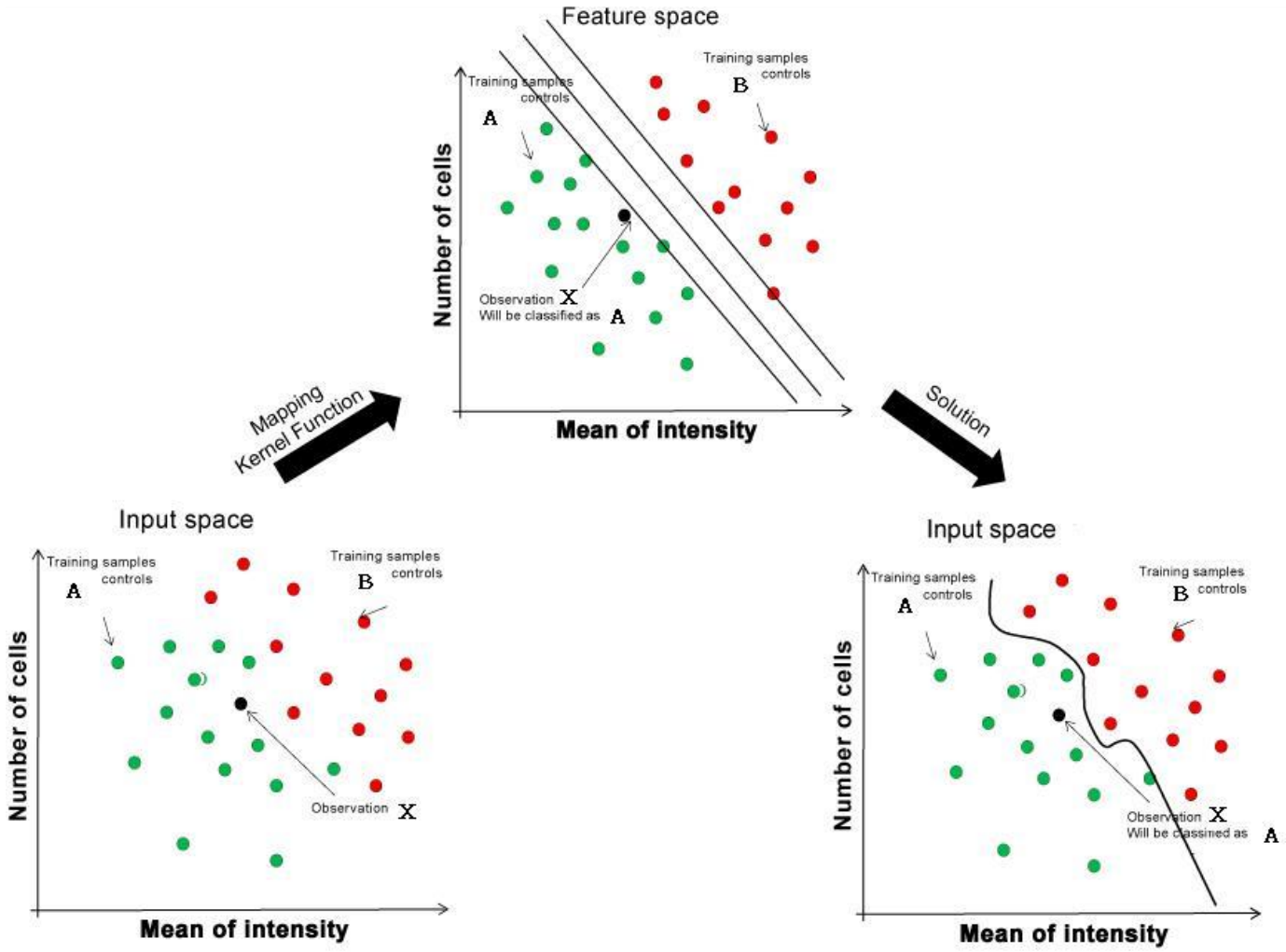
- The task of choosing the most suitable representation is known as feature selection.

- A set of features that describes one case (i.e., a row of predictor values) is called a feature vector.

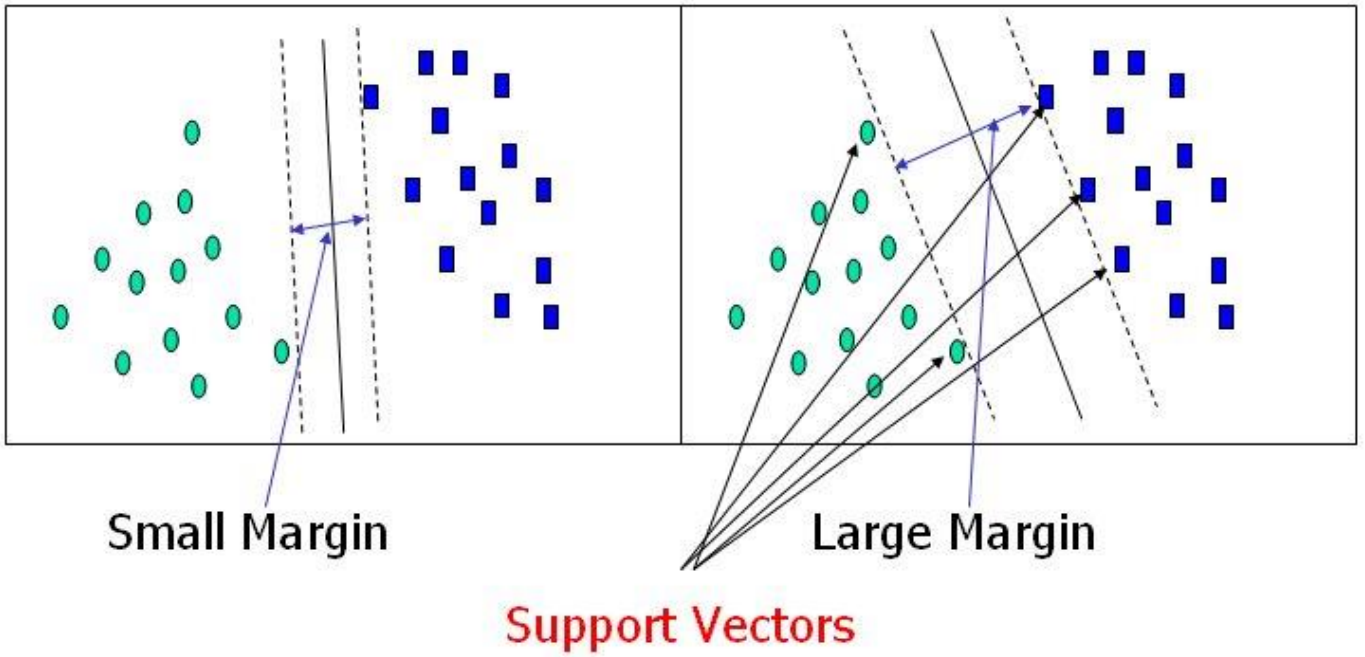
- The goal of SVM modeling is to find the optimal hyperplane that separates clusters of vectors such that the cases with one category of the target variable are on one side of the plane and cases with the other category are on the other side of the plane.

- The vectors near the hyperplane are the support vectors.

Support Vector Machines



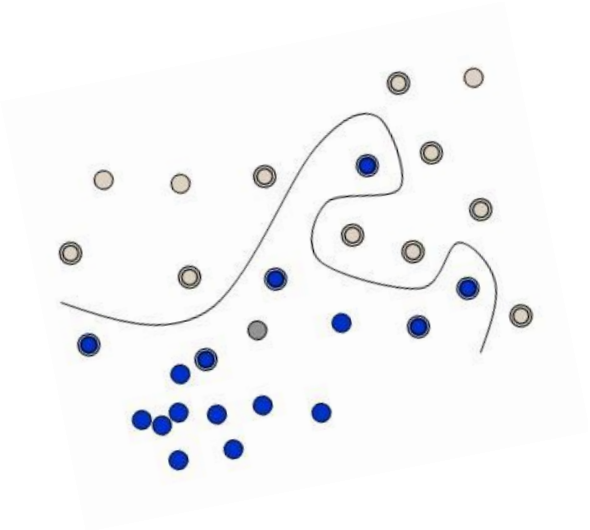
Support Vector Machines

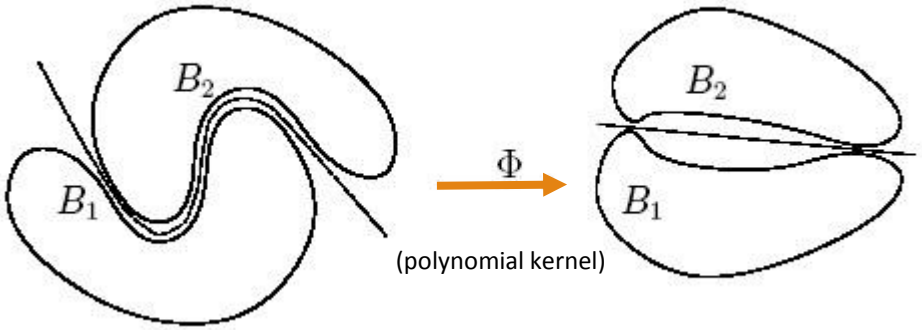


- SVM analysis finds the line (in general, hyperplane) in feature space that is positioned so that the margin between the support vectors is maximized.

- The separating line in the right panel is superior to the line in the left panel
  - Larger margin separating the clusters
  - More support vectors supporting the separation line

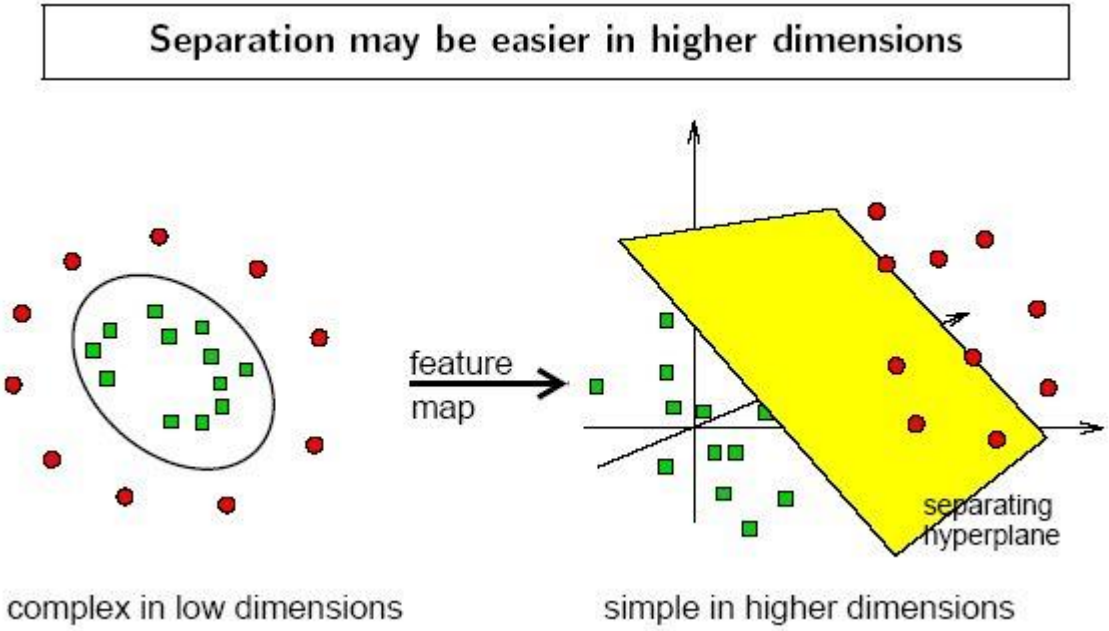
Support Vector Machines





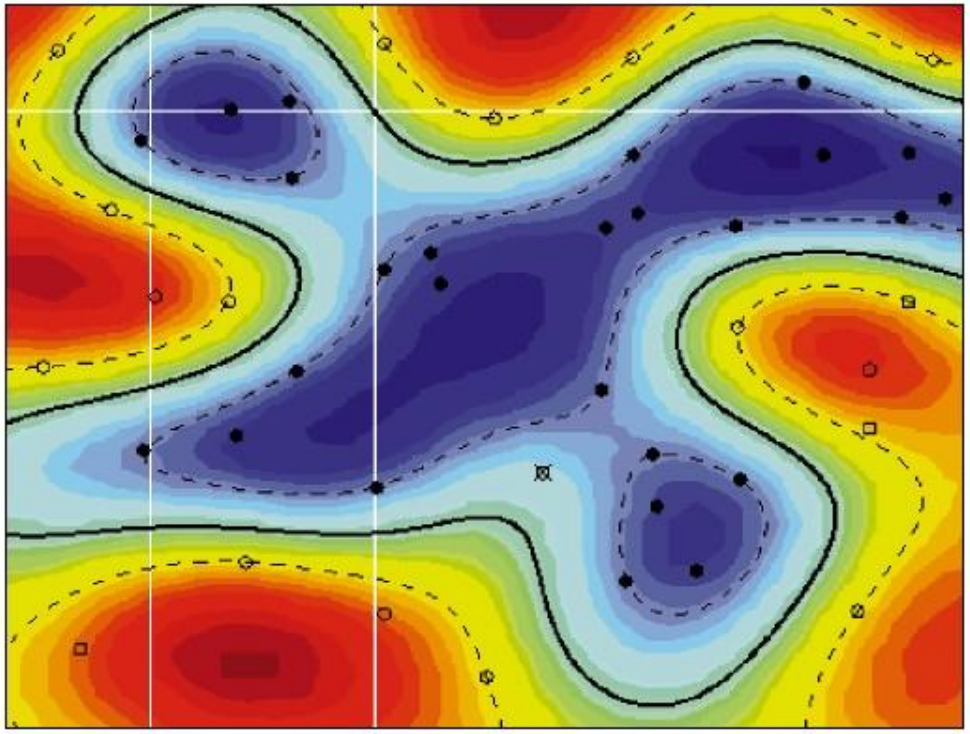
- When a nonlinear dividing line is needed, instead of a fitting nonlinear curves to the data, SVM uses a kernel function **Φ** to transform the data into a higher dimensional space where a hyperplane can be used to perform the separation.
- This separating hyperplane is then mapped back into the input space

Support Vector Machines



- **Using a radial kernel to transform the input space to find the hyperplane in feature space with maximal margin that separates the two classes**

Support Vector Machines

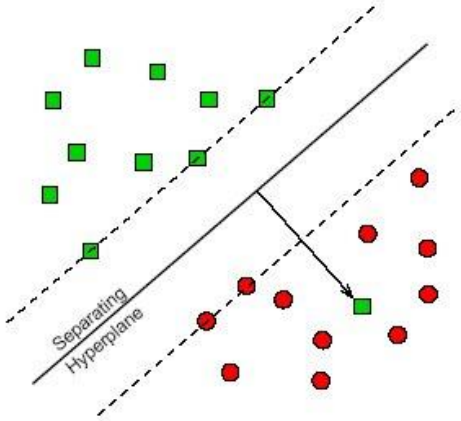


- **Kernel mapping functions allow SVM models to perform accurate separations even with very complex boundaries**

Support Vector Machines

Non-separable training sets

Use linear separation, but admit training errors.



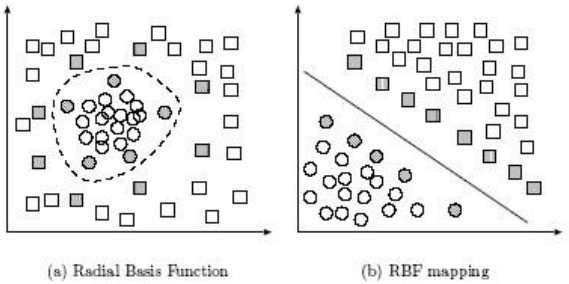
Penalty of error: distance to hyperplane multiplied by *error cost* C.

- Perfect separation is often not possible, or it may result in a model with so many feature vector dimensions that the model overfits the training data and does not generalize well to other data
- SVM models have a cost parameter C that controls the trade off between allowing training errors and forcing rigid margins.
- The cost parameter C creates a soft margin that permits some misclassifications.
- Increasing the value of C increases the cost of misclassifying points and forces the creation of a more accurate model that may not generalize well.

Support Vector Machines

- **Many kernel mapping functions available. However, a few kernel functions have been found to work well for a wide variety of applications:**
  - **Linear**
  - **Radial Basis Function (RBF - default)**
  - **Polynomial**
  - **Sigmoidal**

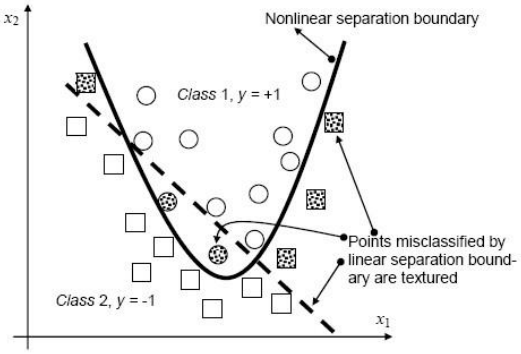
Radial



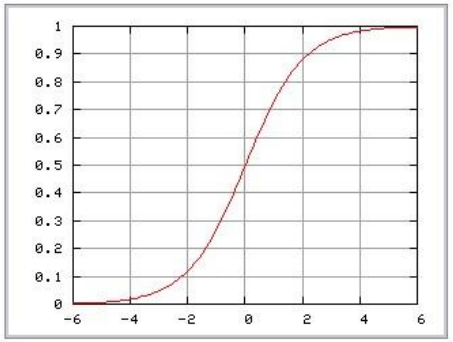
(a) Radial Basis Function (b) RBF mapping

Separable classification with Radial Basis kernel functions in different space. Left: original space. Right: feature space.

Polynomial



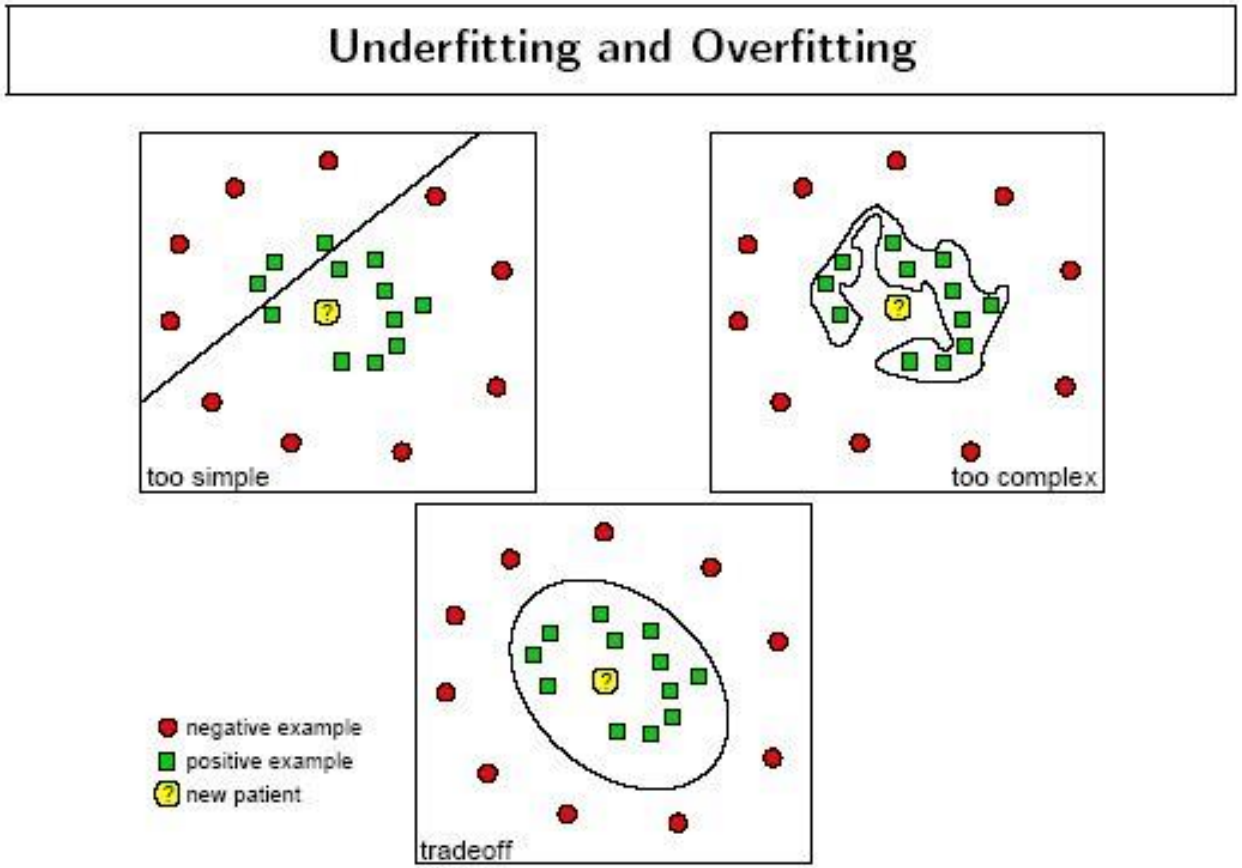
Sigmoidal



Support Vector Machines

- The accuracy of an SVM model is largely dependent on the selection of the kernel and the kernel parameters
- There are different search methods to find the optimal kernel and model parameters
- Finding the optimal cost parameter C is also important to insure that the model doesn't overfit the data, and can generalize well to other datasets

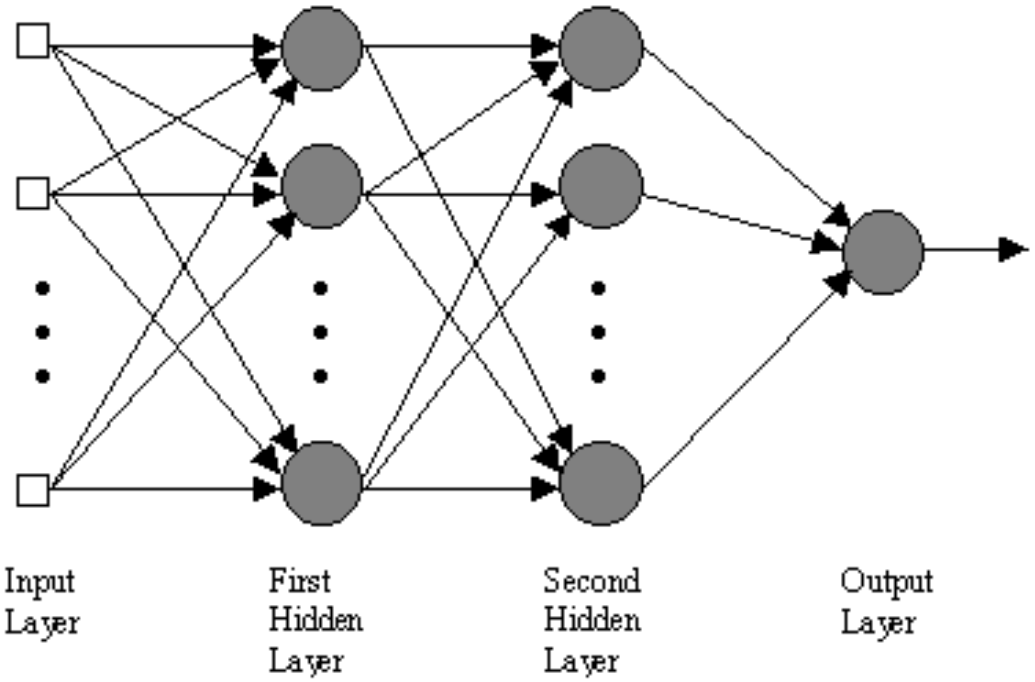
Support Vector Machines



Neural Networks

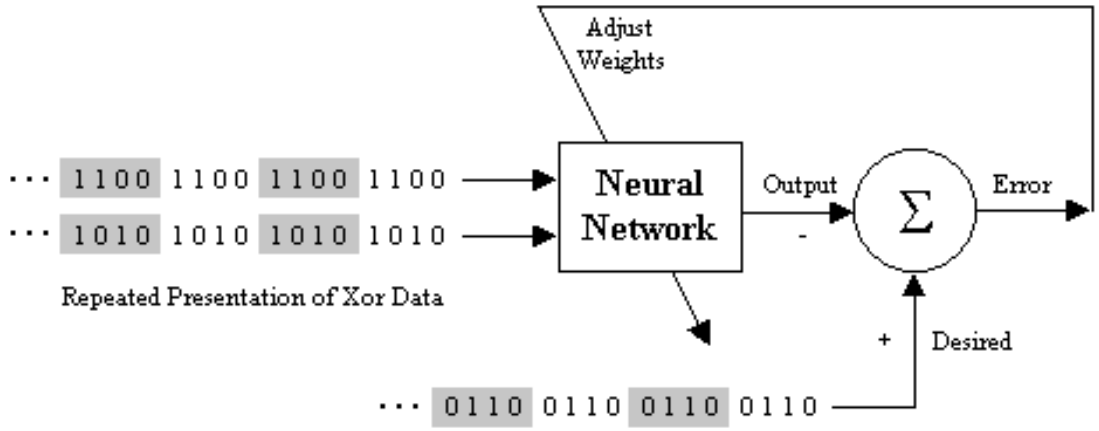
- Designed to capture complex input/output relationships.
- An artificial system designed to make "intelligent" decisions similar to those performed by the human brain.
- Neural networks resemble the human brain in the following two ways:
  - Acquires knowledge through learning.
  - Knowledge is stored within inter-neuron connection strengths known as synaptic weights.
- able to represent both linear and non-linear relationships and can learn these relationships directly from the data being modeled.
- The most common neural network model is the multilayer perceptron (MLP).
  - A supervised-training model designed to correctly map the input to the output using historical data
- The model can have great predictive accuracy, but almost impossible to interpret
- Can be difficult to train – especially as the layers increase

Multilayer Perceptron (MLP)

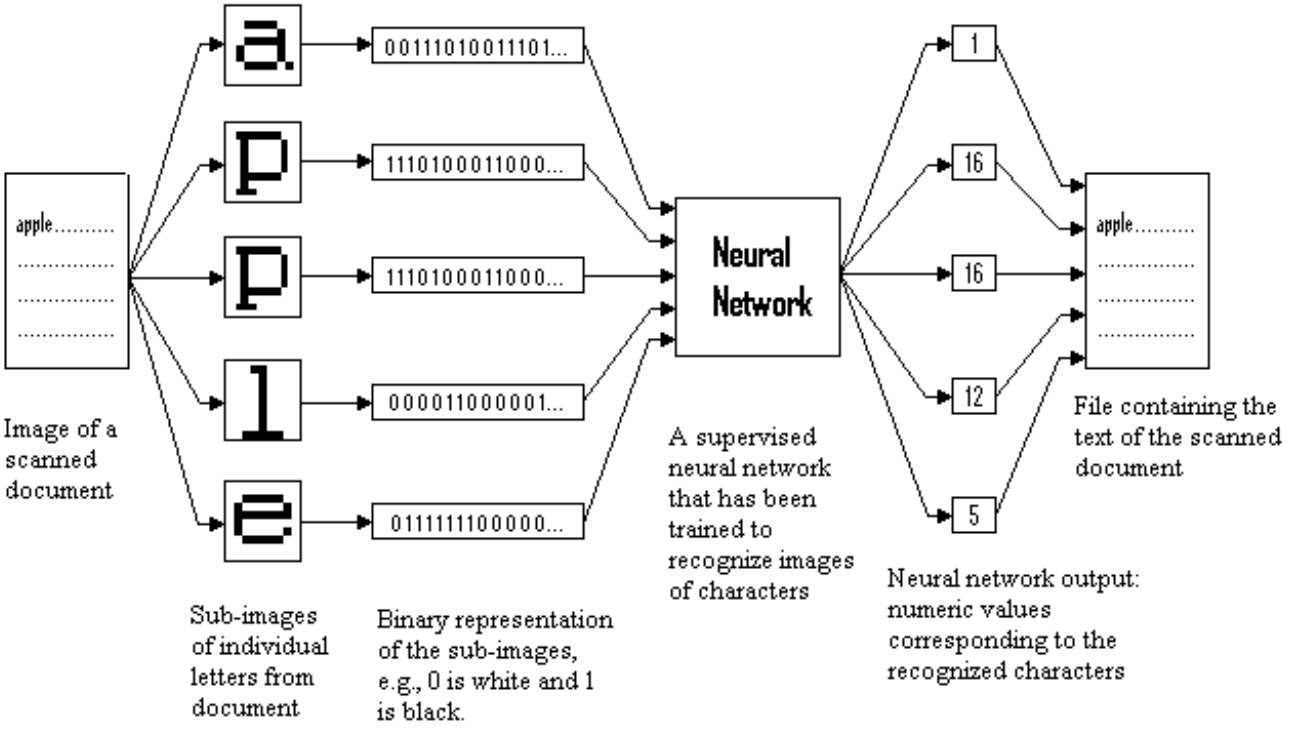


Multilayer Perceptron (MLP)

- The MLP (and many other neural networks) learn using an algorithm called backpropagation.
  - Input data is repeatedly presented to the neural network.
  - With each presentation the output of the neural network is compared to the desired output and an error is computed.
  - This error is then fed back (backpropagated) to the neural network and used to adjust the weights such that the error decreases with each iteration
  - This process is iterated as long as the accuracy in predicting the correct output increases



Application to Optical Character Recognition OCR



- **Also been used to model handwriting recognition, computer vision, speech recognition, protein folding, biomarker discovery, and many others**

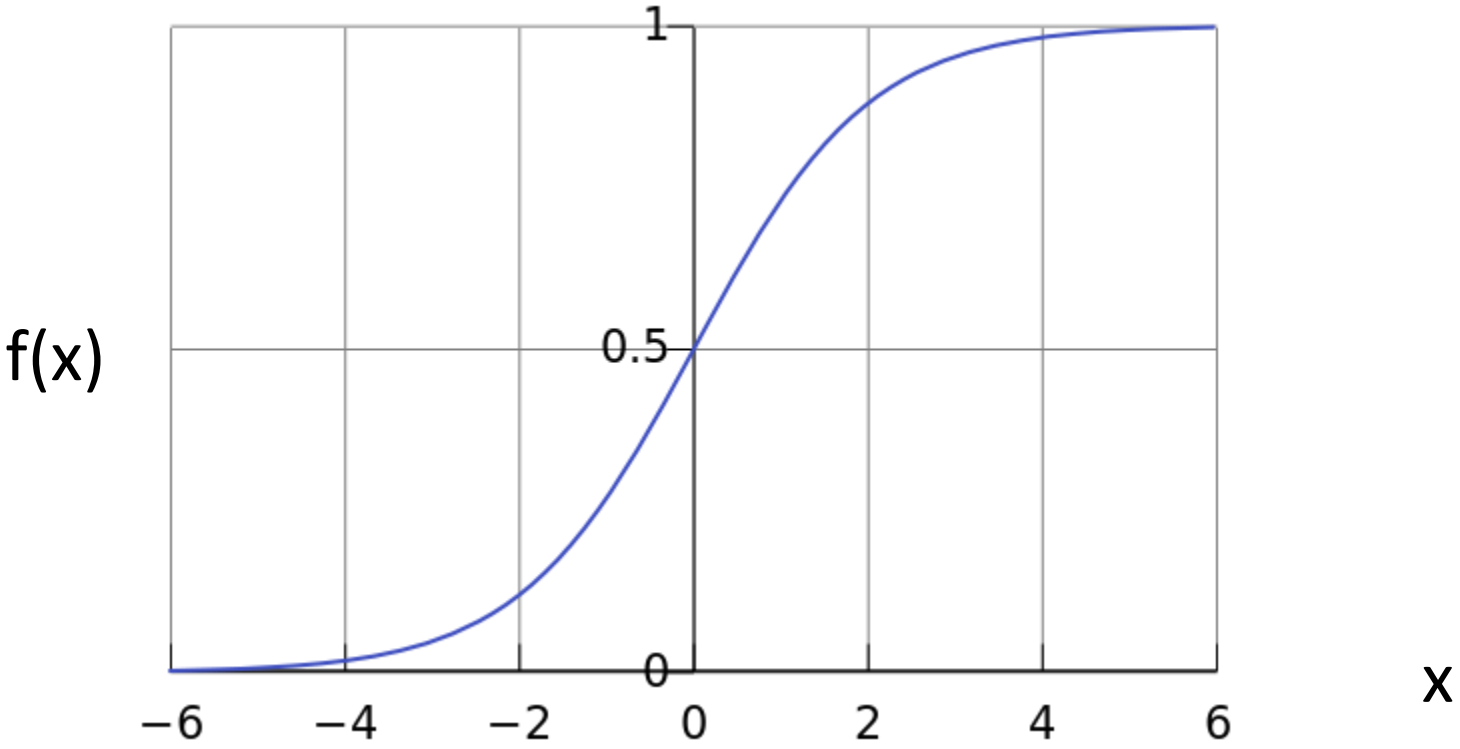
Logistic Regression

- Logistic regression is a form of generalized linear models that has been developed specifically for predicting the value of dependent variables that are binary (categorical with two possible values)
    - such as yes/no, male/female, presence/absence, or diseased/healthy response variables.
    - Independent variables in a logistic regression can be continuous or categorical
    - Can also include interaction effects and weights
- Assumptions of logistic regression models include:
    - independent observations
    - independent variables are linearly related to the log odds (logit) of the dependent variable.
- The dependent variable must be coded as 0 or 1
- The mean of the sample is equal to the proportion of 1s in the sample, and this mean is also the probability of drawing a case labeled as 1 at random from your data.
- The goal is to predict the category for the dependent variable from the suite of independent variables that provides the best balance between model performance and model complexity.
- Rather than directly estimating the category of the response variable at a given value of the predictor, for logistic regression you estimate the probability that an observation has a code of 1. Specifically, logistic regression predicts the log odds that an observation will be a 1 (probability = P) rather than a 0 (probability = 1 - P).

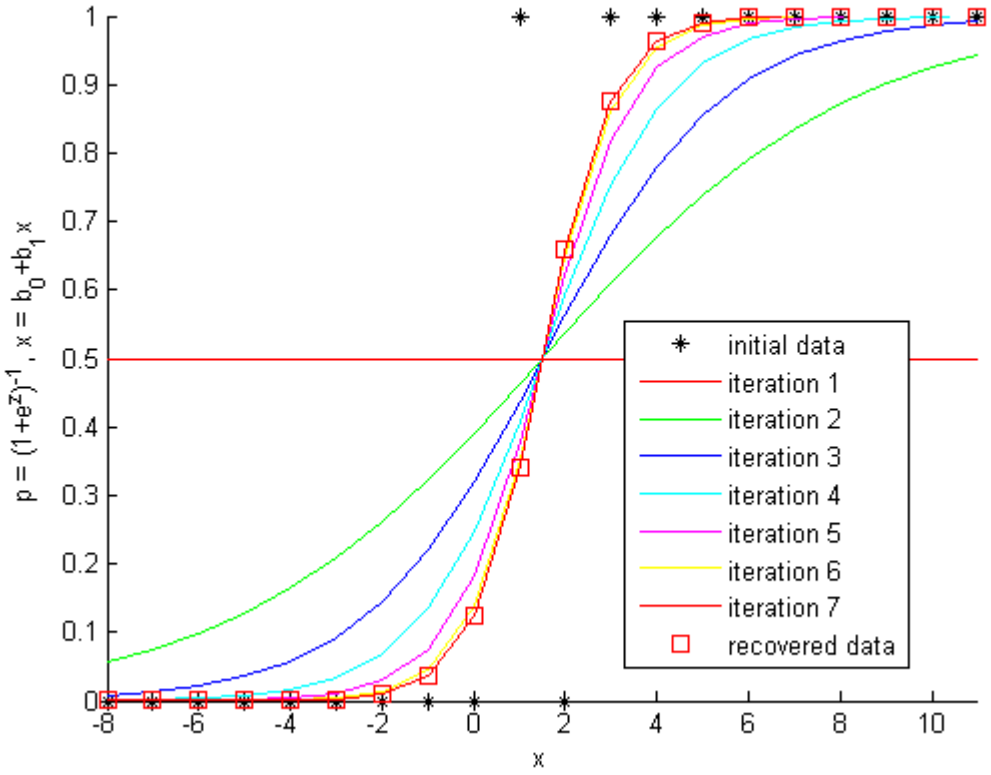
Logistic Curve

- A logistic function or logistic curve is a common special case of the more general sigmoid function, with equation:

$$f(x) = \frac{e^x}{e^x + 1} = \frac{1}{1 + e^{-x}}$$



Logistic Regression



Log odds = Logit Link Linear Model

$$\ln\left(\frac{P}{1-P}\right) = a + bX$$

$$\frac{P}{1-P} = e^{a+bX}$$

$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

Logistic Curve

- **The logit link function constrains the probability values (predictions of y) to a range from 0 to 1**