Support vector machine

- Support vector machine: an approach for classification that was developed in the computer science community in the 90's.
- Hyperplane: in $\mathcal{R}^p$, a hyperplane is a flat subspace of dimension p-1.
- In $\mathcal{R}^2$, a hyperplane is line. In $\mathcal{R}^3$, a hyperplane is a flat two-dimensional subspace: a plane.
- A hyperplane is defined by the equation:

$$b + w_1 x_1 + w_2 x_2 + \cdots + w_p x_p = 0.$$

Support vector machine

- For a point $(x_1, x_2, \cdots, x_p)$, it can either be on the plane:

$$b + w_1 x_1 + w_2 x_2 + \cdots + w_p x_p = 0.$$

  or lie on either side of the plane:

$$b + w_1 x_1 + w_2 x_2 + \cdots + w_p x_p > 0.$$
$$b + w_1 x_1 + w_2 x_2 + \cdots + w_p x_p < 0.$$

- Hyperplane can be used to do classification.

Support vector machine

- Suppose we have a data matrix $X \in \mathcal{R}^{n \times p}$ and labels $y \in [-1, 1]^n$. We also have an observation without a label $x^* = (x_1^*, x_2^*, \cdots, x_p^*)$
- If such a hyperplane that can be used to do classification, what characteristic does it have?

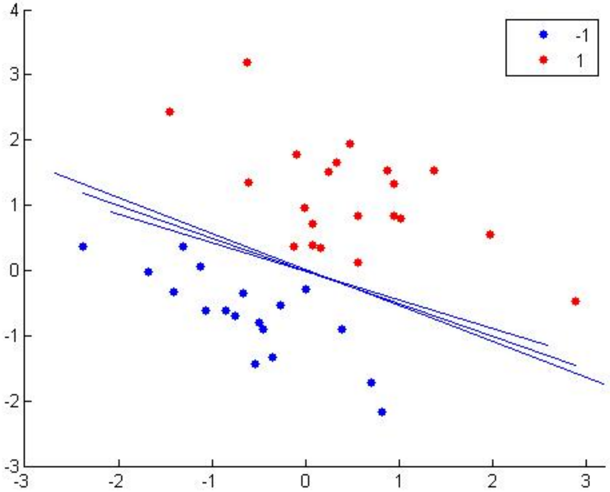
$$b + w_1 x_1 + w_2 x_2 + \cdots + w_p x_p > 0 \text{ if } y_1 = 1.$$
$$b + w_1 x_1 + w_2 x_2 + \cdots + w_p x_p < 0 \text{ if } y_1 = -1.$$

- In general, a separating hyperplane has the property:

$$y_i(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p) > 0$$

for $i = 1, 2, \cdots, p$.

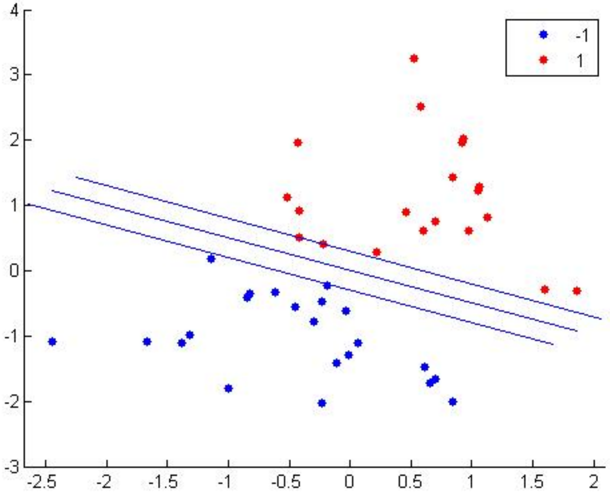
SVM



SVM

- A new observation is assigned to a class depending on which side of the hyperplane it is on.
- If the data is separable, there are infinite number of such separating hyperplanes.
- Maximal margin hyperplane (optimal separating hyper plane) is the separating hyperplane that is farthest from the training observations.
- Given a hyperplane, we can calculate the distance from each observation to the plane. The smallest such distance is called the margin.
- With large margin, we hope that it can reduce the test error on unobserved data.

SVM



SVM formulation

- Consider the case where the two classes are separable. We are looking for a hyperplane defined by $b + w^T x = 0$.
- Also define the class 1 hyperplane as : $b + w^T x = 1$.
- Also define the class -1 hyperplane as : $b + w^T x = -1$.
- What the distance between these two planes?
- Given x_1 on $b + w^T x = 1$, distance between x_1 to $b + w^T x = -1$: $D = \frac{w^T x_1 + b + 1}{\|w\|} = \frac{2}{\|w\|}$.
- We can see that the class 1 or -1 hyperplanes are actually defined by some data points.

SVM formulation

- Looking for a maximal margin classifier will be equivalent to minimizing $\|w\|$.
- Moreover, $+1$ data points have to be on one side of the class 1 hyperplane:

$$\text{If } y_i = 1, w^T x_i + b \geq 1.$$

- -1 data points have to be on one side of the class -1 hyperplane:

$$\text{If } y_i = -1, w^T x_i + b \leq -1.$$

- Those inequalities can be summaried by:

$$y_i(w^T x + b) \geq 1.$$

SVM formulation

- For separable data, the maximal margin classifier is the solution of the following optimization problem:

$$\min_{w,b} \|w\|^2 \text{ s.t: } y_i(w^T x_i + b) \geq 1, i = 1, 2, \cdots, n.$$

- This is a quadratic programming problem.
- Standard form a quadratic programming (QP):

$$\min_{x} \frac{1}{2} x^T Q x + b^T x : Ax = b, Cx \leq d.$$

- Separable SVM in standard QP form:

$$\min_{w,b} \frac{1}{2} w^T \mathbb{I} w : y_i(w^T x_i + b) \geq 1, i = 1, 2, \cdots, n.$$

SVM formulation

- svm_example.1 and svm_separable.m

SVM formulation

- We have the assumption that the data can be perfectly separable into 2 classes. This often does not happen.
- Due to factors such as noise, outliers, error when collecting data, most real data sets are nonseparable.
- The new problem still keeps the maximal margin part, however, it will allow the classifider to make errors.
- How can this be input into the SVM formulation?

SVM formulation

- For an observation, instead of: $y_i(b + w^T x_i) \geq 1$, we let:

$$y_i(b + w^T x_i) \geq 1 - \epsilon_i, \epsilon_i \geq 0.$$

  For example, a class 1 observation is allowed to make an error amount ϵ_i: $b + w^T x_i \geq 1 - \epsilon_i$. For a large enough ϵ_i, the classifier can actually predict this observation to be in -1 class.

- However, we need to control this amount of error, in other words, the amount of error needs to be minimized together with the margin:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \epsilon_i \text{ s.t:}$$

$$y_i(w^T x_i + b) \geq 1 - \epsilon_i, i = 1, 2, \cdots, n., \epsilon_i \geq 0.$$

SVM formulation

- $C > 0$ is a tuning parameter that controls the trade off between the amount of error and the magnitude of the margin.
- It can be shown that the problem above is equivalent to:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} max(0, 1 - y_i(W^T x_i + b)).$$

- Consider the problem above, if we replace $a_i = max(0, 1 - y_i(W^T x_i + b))$, it becomes:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} a_i \text{ s.t:}$$

$$a_i \geq 0, a_i \geq 1 - y_i(W^T x_i + b)), i = 1, 2, \cdots, n.$$

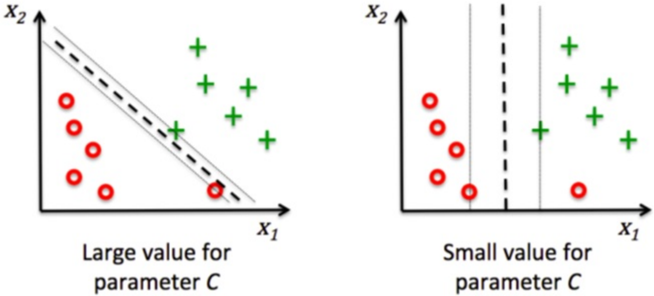
SVM formulation

- $max(0, 1 - y_i(W^T x_i + b))$ is called the hinge loss. It is a convex function
- The nonseparable SVM formulation and the ridge linear regression are similar.
- Both involved minimizing a function that consists of a convex loss function (least square vs. hinge) and a ridge type penalty.
- SVM formulation does not a closed form solution like ridge.

SVM formulation

- Large value of C leads to a classifier that makes very little mistake on the training data.
- Small value of C leads to a classifier that makes more mistake but has a larger margin.
- Bias variance trade off.
- Optimal choice of tuning parameter can be found using cross validation.
- When p is big, SVM still can have the problem with overfitting.

SVM


x_2
x_1
Large value for parameter C
x_2
x_1
Small value for parameter C

SVM formulation

- There are two ways of looking at SVM formulation.
- First as a constrained quadratic programming:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \epsilon_i \text{ s.t:}$$

$$y_i(w^T x_i + b) \geq 1 - \epsilon_i, i = 1, 2, \cdots, n., \epsilon_i \geq 0.$$

  This way, you have $n + p + 1$ unknown varibles: b,w,ϵ, and n constraints to the optimization problem.
- For large data set, not a good idea.

SVM formulation

- Second way, as a non-constrained problem:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} max(0, 1 - y_i(W^T x_i + b)).$$

  There are only $n+1$ uknown variables: w, b. Great.

- Too bad, the hinge loss is non-differentiable. So you can't calculate gradient or Hessian.

SVM

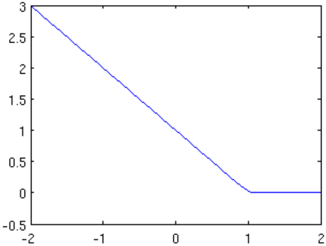


Figure: cvxr.com