Separating Hyperplanes

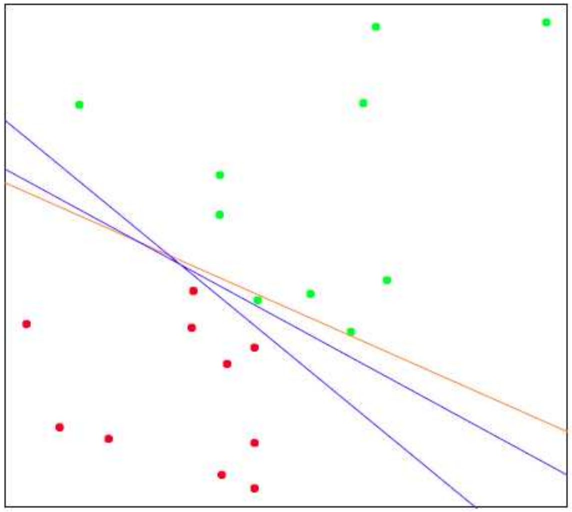


Figure 4.13: *A toy example with two classes separable by a hyperplane. The orange line is the least squares solution, which misclassifies one of the training points. Also shown are two blue separating hyperplanes found by the* perceptron learning algorithm *with different random starts.*

Hyperplanes

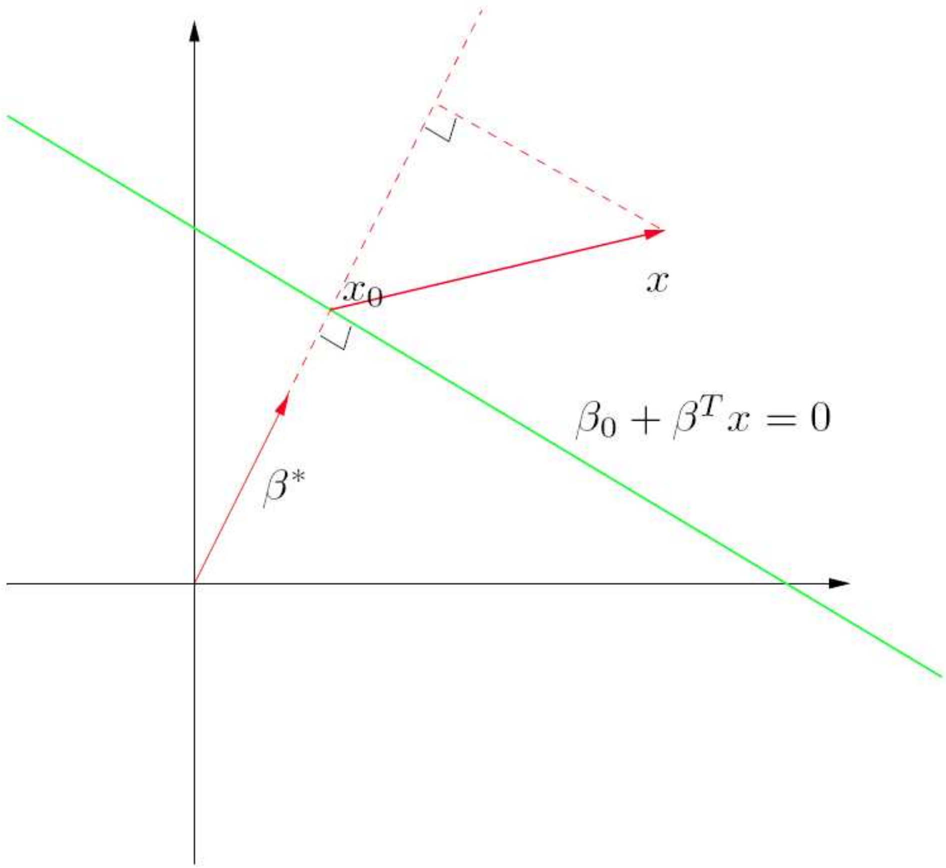


Figure 4.14: *The linear algebra of a hyperplane (affine set).*

Hyperplanes…

Unit vector perpendicular to the hyperplane: $\beta^* = \frac{\beta}{\|\beta\|}$

Signed distance of a point x to the hyperplane: $\frac{\beta^T x + \beta_0}{\|\beta\|}$

Rosenblatt's Perceptron Learning

- Minimize the distance of misclassified points to the decision boundary

$$D(\beta, \beta_0) = -\sum y_i (x_i^T \beta + \beta_0)$$

$$\partial \frac{D(\beta, \beta_0)}{\partial \beta} = -\sum y_i x_i,$$

$$\partial \frac{D(\beta, \beta_0)}{\partial \beta_0} = -\sum y_i.$$

Rosenblatt’s Perceptron Learning…

- Use Stochastic Gradient Descent which does "on-line” updates (take one observation at a time) until convergence:

$$\begin{pmatrix} \beta \\ \beta_0 \end{pmatrix} \leftarrow \begin{pmatrix} \beta \\ \beta_0 \end{pmatrix} + \rho \begin{pmatrix} y_i x_i \\ y_i \end{pmatrix}$$

  - Faster for large data sets
  - Usually ρ=c/#iteration is the learning rate
  - If classes are linearly separable then the process converges in a finite number of steps.
  - neural network which also use SGD

Perceptron Algorithm

$$\begin{pmatrix} \beta \\ \beta_0 \end{pmatrix} \leftarrow 0$$

Keep looping

Choose a point *xi* for which $y_i(\beta_0 + \beta^T x_i) < 0$

$$\begin{pmatrix} \beta \\ \beta_0 \end{pmatrix} \leftarrow \begin{pmatrix} \beta \\ \beta_0 \end{pmatrix} + \begin{pmatrix} y_i x_i \\ y_i \end{pmatrix}$$

Does this algorithm stop?

Convergence Proof

For notational convenience consider the hyperplane equation as: $\beta^T x = 0$
i.e., the input vector x has a leading 1

Because the points are linearly separable there exist a unit vector and a non-negative number such that $y_i \alpha^T x_i \geq \varepsilon > 0, \ \forall i$

Also, we assume that the norm of points are bounded $\|x\| \leq B, \ \forall i$

$$\beta_{k+1}^T \alpha = (\beta_k + y_i x_i)^T \alpha \geq \beta_k^T \alpha + \varepsilon \geq k\varepsilon$$

$$(\beta_{k+1}^T \alpha)^2 \leq \|\beta_{k+1}\|^2 \|\alpha\|^2 \leq \|\beta_{k+1}\|^2 = \|\beta_k + y_i x_i\|^2 = \|\beta_k\|^2 + 2 y_i x_i^T \beta_k + \|y_i x_i\|^2 \leq \|\beta_k\|^2 + \|y_i x_i\|^2 = \|\beta_k\|^2 + B^2 \leq kB^2$$

Cauchy-Swartz inequality

Combining the two inequalities we have $(k\varepsilon)^2 \leq kB^2 \Rightarrow k \leq \left(\frac{B}{\varepsilon}\right)^2$

The iteration number *k* is bounded above, i.e., the algorithm converges

Rosenblatt’s Perceptron Learning

- Criticism
  - Many solutions for separable case
  - SGD converges slowly
  - For non-separable case, it will not converge

Optimal Separating Hyperplane

Consider a linearly separable binary classification problem.

Perceptron training results in one of many possible separating hyperplanes.

Is there an optimal separating hyperplane?

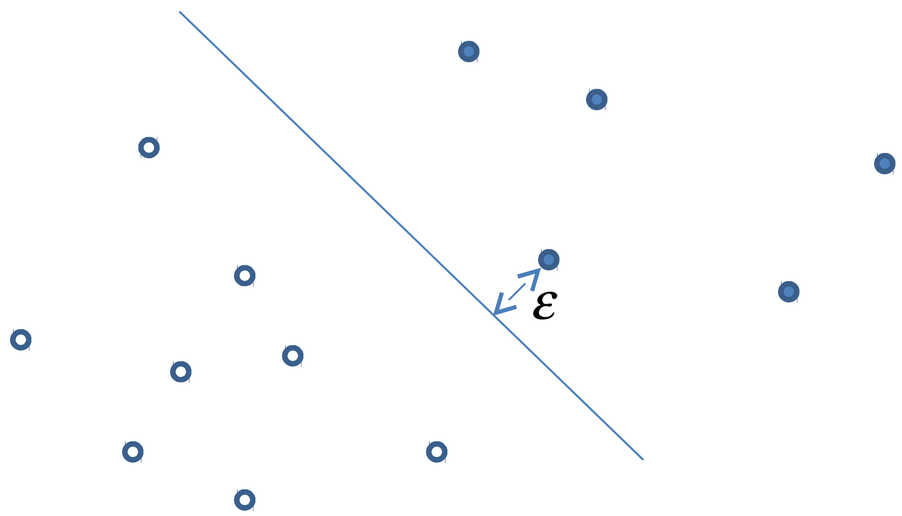
How can we find it out?

Optimal Separating Hyperplane

A quick recall, because of linearly separable classes, we have

$$y_i(\beta_0 + \beta^T x_i) \geq \varepsilon > 0, \ \forall i$$

For a separable hyperplane $\beta_0 + \beta^T x = 0$ with $\|\beta\| = 1,$ the number ε is known as the margin– it is the perpendicular distance to the nearest point from the hyperplane



Why not maximize the margin to obtain the optimal separating hyperplane?

Maximum Margin Hyperplane

$$\max_{\beta_0,\beta} \varepsilon$$

$$\text{subject to}: \quad y_i(\beta_0 + \beta^T x_i) \geq \varepsilon, \ \forall i$$

$$\|\beta\| = 1$$

Equivalent to:

$$\max_{\beta_0,\beta} \frac{\varepsilon}{\|\beta\|}$$

$$\text{subject to}: \quad y_i(\beta_0 + \beta^T x_i) \geq \varepsilon, \ \forall i$$

Equivalent to:

$$\max_{\beta_0,\beta} \frac{1}{\|\beta\|}$$

$$\text{subject to}: \quad y_i(\beta_0 + \beta^T x_i) \geq 1, \ \forall i$$

Equivalent to:

$$\min_{\beta_0,\beta} \frac{1}{2}\|\beta\|^2$$

$$\text{subject to}: \quad y_i(\beta_0 + \beta^T x_i) \geq 1, \ \forall i$$

Convex quadratic programming

$\Longrightarrow$ unique solution

Solving Convex QP: Lagrangian

Lagrangian function: $L = \frac{1}{2}\|\beta\|^2 - \sum_{i=1}^{N} \alpha_i [y_i (x_i^T \beta + \beta_0) - 1]$

α_i's are non-negative Lagrangian multipliers

Why do Lagrangian multiplers exist for this optimization problem?

A remarkable property of linear constraints is that it always guarantees the existence of Lagrange multipliers when a (local) minimum of the optimization problem exists (see D.P. Bertsekas, *Nonlinear programming*)

Also see Slater constraint qualification to know about the existence of Lagrange Multipliers (see D.P. Bertsekas, *Nonlinear programming*)

Lagrangian function plays a central role in solving the QP here

Solving Convex QP: Duality

$$\min_{\beta_0,\beta} \frac{1}{2}\|\beta\|^2$$

$$\text{subject to:} \qquad y_i(\beta_0 + \beta^T x_i) \geq 1, \ \forall i$$

Primal optimization problem

$$L = \frac{1}{2}\|\beta\|^2 - \sum_{i=1}^{N} \alpha_i [y_i(x_i^T \beta + \beta_0) - 1],$$

$$\alpha_i \geq 0, \ \forall i.$$

Lagrangian function

$$q(\alpha) = \min_{\beta,\beta_0} L = \min_{\beta,\beta_0} \{\frac{1}{2}\|\beta\|^2 - \sum_{i=1}^{N} \alpha_i [y_i(x_i^T \beta + \beta_0) - 1]\}$$

Dual function

$$\max_{\alpha} q(\alpha)$$

$$\text{subject to:} \qquad \alpha_i \geq 0, \ \forall i$$

Dual optimization problem
(here a simpler optimization problem)

Solving Convex QP: Duality…

A remarkable property of convex QP with linear inequality constraints
s that there is no duality gap

This means the solution value of the primal and the dual problems are same

The right strategy here is to solve the dual optimization problem
a simpler problem) and obtain corresponding the primal problem solution

We will learn about another important reason (kernel) for solving the dual proble

Finding Dual Function

- Lagrangian function minimization

$$L = \frac{1}{2}\|\beta\|^2 - \sum_{i=1}^{N} \alpha_i [y_i (x_i^T \beta + \beta_0) - 1]$$

- Solve:

$$\frac{\partial L}{\partial \beta} = \beta - \sum_i \alpha_i y_i x_i = 0 \qquad (1)$$

$$\frac{\partial L}{\partial \beta_0} = \sum_i \alpha_i y_i = 0 \qquad (2)$$

- Substitute (1) and (2) in L to form the dual function:

$$q = \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_k \alpha_i \alpha_k y_i y_k x_i^T x_k$$

$$\text{subject to}: (2)$$

Dual Function Optimization

$$\max_{\alpha_1,\ldots,\alpha_N} \{\sum_i \alpha_i - \frac{1}{2}\sum_i \sum_k \alpha_i \alpha_k y_i y_k x_i^T x_k\}$$

$$\text{subject to}: \alpha_i \geq 0, \forall i,$$

$$\sum_i \alpha_i y_i = 0.$$

Dual problem

Equivalent to:

$$\min_{\alpha_1,\ldots,\alpha_N} \{\frac{1}{2}\sum_i \sum_k \alpha_i \alpha_k y_i y_k x_i^T x_k - \sum_i \alpha_i\}$$

$$\text{subject to}: \quad \alpha_i \geq 0, \forall i,$$

$$\sum_i \alpha_i y_i = 0.$$

Dual problem
(simpler optimization)

In matrix vector form

$$\frac{1}{2}\alpha^T [diag(y) X X^Y diag(y)] \alpha - 1^T \alpha,$$

$$\text{subject to}: \quad \alpha \geq 0,$$

$$y^T \alpha = 0.$$

⟺ Compare the implementation simple_svm.m

Optimal Hyperplane

After solving the dual problem we obtain αi ‘s;
how do construct the hyperplane from here?

To obtain β use the equation: $$\beta = \sum_i \alpha_i y_i x_i$$

How do we obtain $\beta 0$?
We need the complementary slackness criteria, which are the results of Karush-Kuhn-Tucker (KKT) conditions for the primal optimization problem.

Complementary slackness means:

$$\alpha_i > 0 \Rightarrow y_i(x_i^T \beta + \beta_0) = 1,$$
$$y_i(x_i^T \beta + \beta_0) > 1 \Rightarrow \alpha_i = 0.$$

Training points corresponding to non-negative αi ‘s are support vectors.

$\beta 0$ is computed from $y_i(x_i^T \beta + \beta_0) = 1$ for which αi ‘s are non-negative.

Optimal Hyperplane/Support Vector Classifier

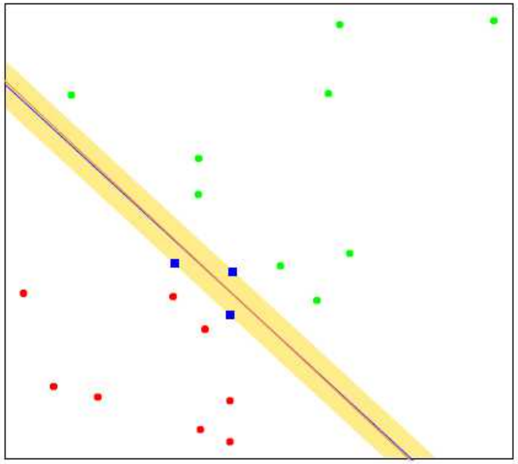


Figure 4.15: *The same data as in Figure 4.13. The shaded region delineates the maximum margin separating the two classes. There are three support points indicated, which lie on the boundary of the margin, and the optimal separating hyperplane (blue line) bisects the slab. Included in the figure is the boundary found using logistic regression (red line), which is very close to the optimal separating hyperplane (see Section 12.3.3).*

In interesting interpretation from the equality constraint in the dual problem is as follows.

$$\sum_i \alpha_i y_i = 0 \Rightarrow$$

αi are forces on both sides of the hyperplane, and the net force is zero on the hyperplane.

Karush-Kuhn-Tucker Conditions

- Karush-Kuhn-Tucker (KKT) conditions
  - A generalization of Lagrange multipliers, for inequality constraints

$$\min_x f(x) \quad \text{subject to}$$

$$\mathrm{g_i}(x) \le 0 (i = 1, \ldots, m),$$

$$\mathrm{h_j}(x) = 0 (j = 1, \ldots, l)$$

Optimal Separating Hyperplanes

- Karush-Kuhn-Tucker conditions (KKT)
- Assume $\mathrm{f(x), g_i(x), and\ h_j(x)}$ are convex
- If there exist
  - feasible point x^*
  - $\mu_i \geq 0 (i = 1, \ldots, m)$ and $v_j \geq 0 (j = 1, \ldots, l)$
  - $f'(x^*) + \sum_i \mu_i g_i'(x^*) + \sum_j v_j h_j'(x^*) = 0$
  - s.t. $\mu_i g_i(x^*) = 0, i = 1, \ldots, m$
- then the point x^* is a global minimum.