Introduction to Data Mining

Support Vector Machine

Jun Huang

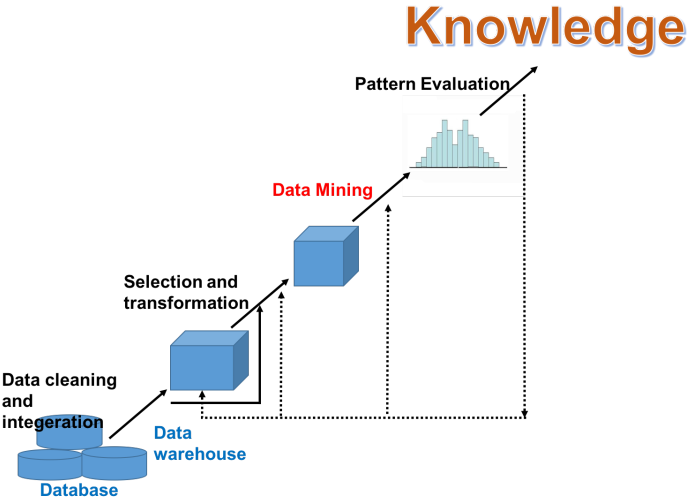
Anhui University of Technology

Spring 2018

huangjun_cs@163.com

KDD Process

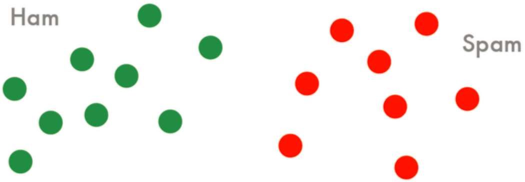
Data Mining-Core of Knowledge discovery process



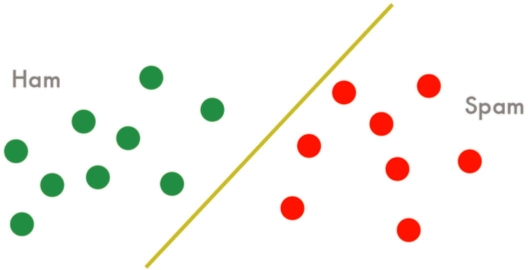
Support Vector Machine: History

- Linear SVM: Cortes and Vapnik, Support Vector Networks, Machine Learning, 1995.
- Kernelized SVM: Boster, Guyon, and Vapnik, A Training Algorithm for Optimal Margin Classifiers, workshop in COLT, 1992.
- SVR: Drucker, burges, Smlola, and Vapnik, Support Vector Regression Machine, NIPS 1996.
- Generalization Analysis: Vapnik, the nature of Statistical Learning Theory, Spinger, 1995.
- Generalization Analysis: Vapnik, Statistical Learning Theory, Wiley&Sons, 1998.
- SMO: Platt, Fast Training of Support Vector Machines Using Sequential Minimal Optimization.
- SVM Light: Joachims, http://svmlight.joachims.org/
- LIBSVM: Chang and Lin, http://www.csie.ntu.edu.tw/ cjlin/libsvm Multi-Class SVM and StructSVM

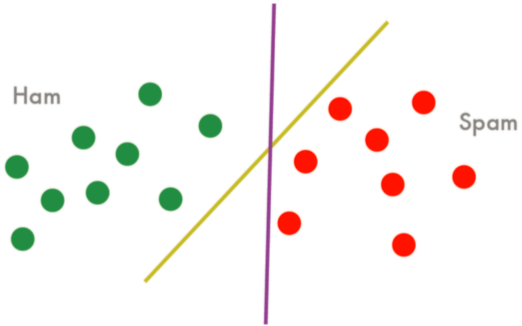
Linearly Separable Classification



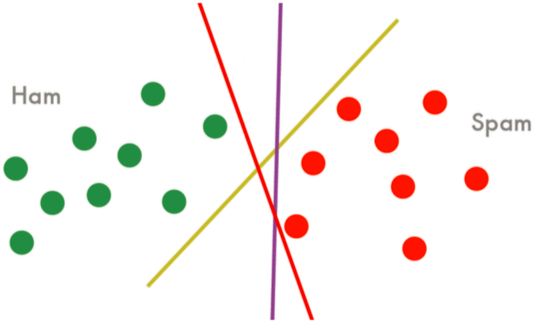
Linearly Separable Classification



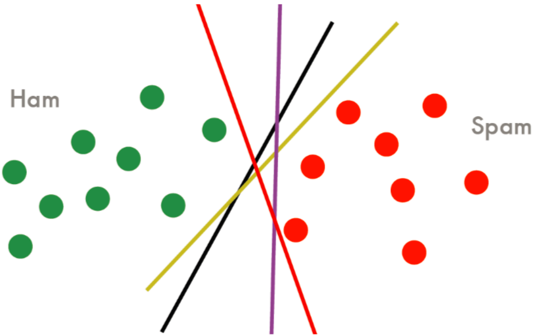
Linearly Separable Classification



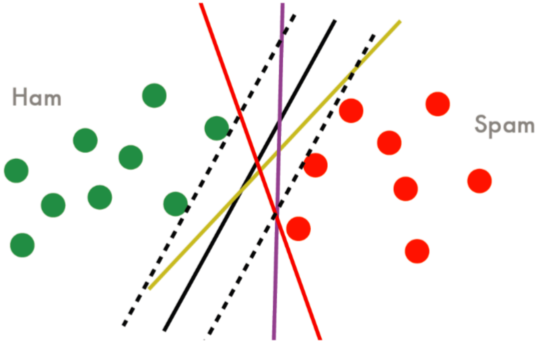
Linearly Separable Classification



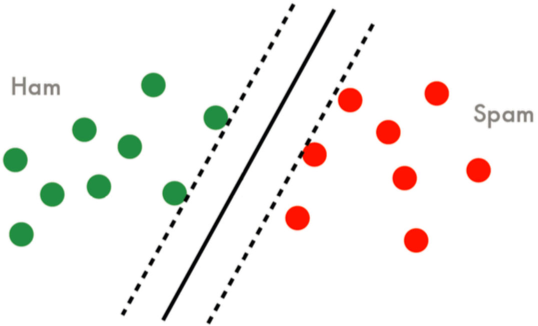
Linearly Separable Classification



Linearly Separable Classification



Optimal Classifier



Margin: Intuition

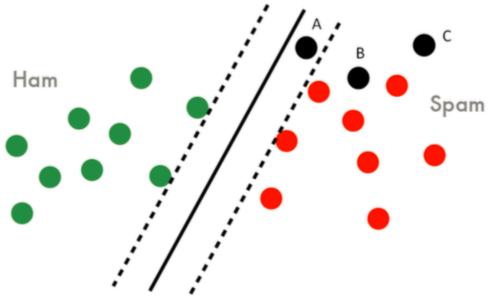
- Logistic Regression

$$P(y=1|x) = \frac{1}{1+e^{-w^T x}}$$

- Predict 1 if $P(y=1|x) \geq 0.5$, i.e. $w^T x \geq 0$;
- Predict 0 if $P(y=-1|x) < 0.5$, i.e. $w^T x < 0$;
- The larger $w^T x$ is, the higher our degree of " confidence" that the label is 1.
- Given a training data, we have found a good fit to the training data if we can find w so that $w^T x \gg 0$ whenever $y=1$, and $w^T x \ll 0$ whenever $y=-1$.

Problem: How to define the confidence of a classifier without probability?

Margin: Intuition



- The confidence of predicting A, B and C to 1 is high, medium, and low, respectively.
- The distance of a point to the separating hyperplane reflects the degree of confidence of prediction.

Notations

- Linear classifier for a binary classification problem with label y and features x.
- $y \in \{-1, 1\}$.
- linear classifier:

$$f_{w,b}(x) = g(w^T x + b)$$

- $g(z) = 1$ if $z \geq 0$, and $g(z) = -1$ otherwise.
- **Hyperplane**: $w^T x + b = 0$.

Functional Margin

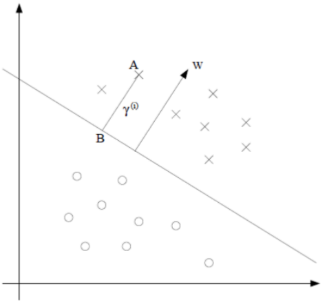
- **Functional margin** of hyperplane (w, b) with respect to the training example $(x^{(i)}, y^{(i)})$:

$$\hat{\gamma}^{(i)} = y^{(i)}(w^T x^{(i)} + b)$$

  - If $y^{(i)} = 1$, we need $w^T x^{(i)} + b$ to be a large positive number.
  - If $y^{(i)} = -1$, we need $w^T x^{(i)} + b$ to be a large negative number.
  - If $y^{(i)}(w^T x^{(i)} + b) > 0$, then our prediction on this example is correct.
  - A large functional margin represents a confident and a correct prediction.
- Given a training set $S = \{(x^{(i)}, y^{(i)})\}_{i=1}^n$, the **functional margin** of (w, b) with respect to S:

$$\hat{\gamma} = \min_{i=1,\cdots,n} \hat{\gamma}^{(i)}$$

Geometric Margin



- For point A, which represents the input $x^{(i)}$ with label $y^{(i)} = 1$, its distance to the decision boundary, $\gamma^{(i)}$, is given by the line segment AB.
- **Question: how to find the value of $\gamma^{(i)}$?**

Geometric Margin

- Point B is given by $x^{(i)} - \gamma^{(i)} w / \|w\|_2$
- B lies in the decision boundary:

$$w^T(x^{(i)} - \gamma^{(i)} \frac{w}{\|w\|_2}) + b = 0$$

- Solving for $\gamma^{(i)}$ yields:

$$\gamma^{(i)} = \frac{w^T(x^{(i)} + b}{\|w\|_2} = (\frac{w}{\|w\|_2})^T x^{(i)} + \frac{b}{\|w\|_2}$$

- The geometric margin of (w, b) with respect to a training example $(x^{(i)}, y^{(i)})$ is:

$$\gamma^{(i)} = y^{(i)} \left((\frac{w}{\|w\|_2})^T x^{(i)} + \frac{b}{\|w\|_2} \right)$$

Geometric Margin

- Given a training set $S = \{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$, the **geometric margin** of (w, b) with respect to S:

$$\gamma = \min_{i=1,2,\ldots,n} \gamma^{(i)}$$

- The relationships between functional margin and geometric margin:

$$\gamma^{(i)} = \frac{\hat{\gamma}^{(i)}}{\|w\|_2}, \quad \gamma = \frac{\hat{\gamma}}{\|w\|_2}$$

- If $\|w\|_2 = 1$, then the functional margin equals the geometric margin
- Invariance property of geometric margin

Optimal Margin Classifier

- Philosophy: Given a training set, a natural desideratum is to try to find a decision boundary that maximizes the (geometric) margin, since this would reflect a very confident set of predictions on the training set and a good "fit" to the training data.
- Assume that we are given a training set that is linearly separable.

$$\max_{\gamma,w,b} \gamma$$

$$s.t. \quad y^{(i)}\left(\frac{w^T}{\|w\|_2}x^{(i)} + \frac{b}{\|w\|_2}\right) \geq \gamma, i = 1, 2, ..., n$$

- Max Margin: We want to maximize γ, subject to each training example having geometric margin at least γ

Optimal Margin Classifier

- First Transforming: (**Key Idea**: Relationships between functional margin and geometric margin)

$$\max_{\hat{\gamma},w,b} \frac{\hat{\gamma}}{\|w\|_2}$$

$$s.t. \quad y^{(i)}\left(\frac{w^T}{\|w\|_2}x^{(i)} + \frac{b}{\|w\|_2}\right) \geq \hat{\gamma}, i = 1, 2, ..., n$$

- Second Transforming: (**Key Idea:** We can add an arbitrary scaling constraint on w and b without changing anything, $s.t. \quad \gamma = 1$)

$$\max_{\hat{\gamma},w,b} \frac{1}{2}\|w\|_2^2$$

$$s.t. \quad y^{(i)}\left(\frac{w^T}{\|w\|_2}x^{(i)} + \frac{b}{\|w\|_2}\right) \geq 1, i = 1, 2, ..., n$$

- Optimization problem with a convex quadratic objective and only linear constraints
- Sovled with commercial quadratic programming(QP) code

Linearly Seperable SVM

- **Input**: a linearly separable training set $S = \{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$
- **output**: a separating hyperplane and decision function
  - Solving the following optimization problem to obtain the optimal classifier (w^*, b^*):

$$\max_{\hat{\gamma}, w, b} \frac{1}{2}\|w\|_2^2$$

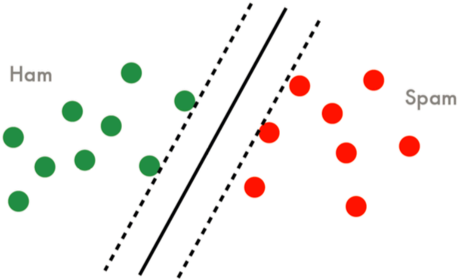
$$s.t. \quad y^{(i)}\left(\frac{w^T}{\|w\|_2}x^{(i)} + \frac{b}{\|w\|_2}\right) \geq 1, i = 1, 2, ..., n$$

  - Separating hyperplane: ${w^*}^T x + b^* = 0$, and decision fucntion:

$$f_{w,b}(x) = sign({w^*}^T x + b^*)$$

Support Vectors and Margin

- **Support Vectors:** the training examples which lie nearest to the separating hyperplane
- For all support vectors, we have: $y(w^T x + b) = 1$



- Functional Margin: $\hat{\gamma} = 1$
- Geometric Margin: $\gamma = \frac{1}{\|w\|_2}$
- Margin: $\frac{2}{\|w\|_2}$

Lagrange Duality: Simplest Case

- Equality constrained optimization problem:

$$\min_{w} f(w)$$
$$s.t. \quad h_i(w) = 0, i = 1, 2, ..., l$$

- Lagrangian:

$$L(w, b) = f(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

- The $\beta_i, i = 1, ..., l$ are called the Lagrange Multipliers
- We would then find and set L's partial derivatives to zero:

$$\frac{\partial L}{\partial w_i} = 0, \frac{\partial L}{\partial \beta_i} = 0$$

Lagrange Duality: Generalized Case

- Optimization problem with both inequality and equality constraints:

$$\begin{aligned} \min_{w} \quad & f(w) \\ s.t. \quad & g_i(w) = 0, i = 1, 2, ..., k \\ & h_i(w) \leq 0, i = 1, 2, ..., l \end{aligned}$$

  - Primal optimization problem
- Generalized Lagrangian:

$$L(w, b) = f(w) + \sum_{i=1}^{l} \alpha_i g_i(w) + \sum_{i=1}^{k} \beta_i h_i(w)$$

- The α_i's and β_i's are the Lagrange Multipliers

Lagrange Duality

- Consider the quantity:

$$\theta_P(w) = \max_{\alpha,\beta:\alpha_i \geq 0} L(w, \alpha, \beta)$$

- If w violates any of the primal constraints, then

$$\theta_P(w) = \max_{\alpha,\beta:\alpha_i \geq 0} f(w) + \sum_{i=1}^{l} \alpha_i h_i(w) + \sum_{i=1}^{k} \beta_i h_i(w) = \infty$$

- If the constraints are indeed statified, then

$$\theta_P(w) = f(w)$$

Lagrange Duality

- Consider the minimization problem:

$$\min \theta_P(w) = \min_{w} \max_{\alpha,\beta:\alpha_i \geq 0} L(w, \alpha, \beta)$$

- This problem has the same solution as our primal problem
- Define the optimal value of the objective to be $p^* = \min_w \theta_P(w)$
- It is called the value of the primal problem

Lagrange Duality

- Consider a different problem:

$$\min \theta_D(\alpha, \beta) = \min_w L(w, \alpha, \beta)$$

- Dual optimization problem:

$$\max_{\alpha,\beta:\alpha_i \geq 0} \theta_D(\alpha, \beta) = \max_{\alpha,\beta:\alpha_i \geq 0} \min_w L(w, \alpha, \beta)$$

- Define the optimal value of the dual problems's objective to be $d^* = \max_{\alpha,\beta:\alpha_i \geq 0} \theta_D(\alpha, \beta)$

Relationships between the primal and the dual problems:

$$d^* = \max_{\alpha,\beta:\alpha_i \geq 0} \min_w L(w, \alpha, \beta) \leq \min_w \max_{\alpha,\beta:\alpha_i \geq 0} L(w, \alpha, \beta) = p^*$$

Lagrange Duality: When will we have $d^* = p^*$?

- Suppose f and g_is are convex, the h_i s are affine, and the constraints g_is are (strictly) feasible.
- There must exists w^*, α^*, β^* so that w^* is the solution to the primal problem, α^*, β^* are the solution to the dual problem, and moreover $p^* = d^* = L(w^*, \alpha^*, \beta^*)$
- w^*, α^*, β^* satisfy the **Karush-Kuhn-Tucker (KKT)** conditions:

$$\frac{\partial L(w^*, \alpha^*, \beta^*)}{\partial w_i} = 0, i = 1, ..., M$$

$$\frac{\partial L(w^*, \alpha^*, \beta^*)}{\partial \beta_i} = 0, i = 1, ..., l$$

$$\alpha_i^* g_i(w^*) = 0, i = 1, ..., k$$

$$g_i(w^*) \leq 0, i = 1, ..., k$$

$$\alpha_i^* \geq 0, i = 1, ..., k$$

- If some w^*, α^*, β^* satisfy the KKT condition, then it is also a solution to the primal and dual problems

Optimal Margin Classifier: Dual Solution

- Prime optimization problem for finding the optimal margin classifier:

$$\min_{w,b} \frac{1}{2}\|w\|_2^2$$
$$s.t. \quad y^{(i)}(w^T x^{(i)} + b) \geq 1, i = 1, ..., n$$

- The constraint can be writen as:

$$g_i(w) = -y^{(i)}(w^T x^{(i)} + b) + 1 \leq 0, i = 1, ..., n$$

- From the KKT dual complementarity condition, we will have $\alpha_i > 0$ only for the training examples that have functional margin exactly equal to 1 (i.e.,$g_i(w) = 0$). These points are called **Support Vectors**
- The member of support vectors can be much smaller than the size of the training set

Optimal Margin Classifier: Dual Solution

- The lagrangian for the optimization problem:

$$L(w, b, \alpha) = \frac{1}{2}\|w\|_2^2 - \sum_{i=1}^{n} \alpha_i [y^{(i)} (w^T x^{(i)} + b) - 1]$$

- Dual solution of the problem
  - Minimize $L(w, b, \alpha)$ with respect to w and b (for fixed α) to get $\theta_D(\alpha)$
  - Find d^* by $\max_\alpha \theta_D(\alpha)$

Optimal Margin Classifier: Dual Solution

- Minimize $L(w, b, \alpha)$ with respect to w and b to get $\theta_D(\alpha)$
  - Setting the derivatives of L with respect to w and b to zero:

$$\nabla_w(w, b, \alpha) = w - \sum_{i=1}^{n} \alpha_i y^{(i)} x^{(i)} = 0$$

  - This implies that $w = \sum_{i=1}^{n} \alpha_i y^{(i)} x^{(i)}$
  - For the derivative with respect to b, we obtain:

$$\frac{\partial L(w, b, \alpha)}{\partial b} = \sum_{i=1}^{n} \alpha_i y^{(i)} = 0$$

  - Plug them back into the lagrangian, we get:

$$\theta_D(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \alpha_i y^{(i)} y^{(j)} (x^{(i)})^T x^{(j)}$$

Optimal Margin Classifier: Dual Solution

- Dual optimization problem:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \alpha_i y^{(i)} y^{(j)} \left(x^{(i)}\right)^T x^{(j)}$$

$$s.t. \quad \alpha_i \geq 0, i = 1, \ldots, n$$

$$\sum_{i=1}^{n} \alpha_i y^{(i)} = 0$$

Optimal Margin Classifier: Primal Solution via Dual Optimization

Theorem (Solution of the Primal Optimization Problem)

Suppose that $\alpha^* = (\alpha_1^*, ..., \alpha_n^*)$ *are the optimal solution of the dual optimization problem, then there exists* j *such that* $\alpha_j^* > 0$, *and we have,*

$$w^* = \sum_{i=1}^{n} \alpha_i^* y^{(i)} x^{(i)}, \quad b^* = y^{(j)} - \sum_{i=1}^{n} \alpha_i^* y^{(i)} \cancel{y^{(j)}} (x^{(i)})^T x^{(j)}$$

- Separating hyperplane:

$$\sum_{i=1}^{n} \alpha_i^* y^{(i)} (x^{(i)})^T x + b^* = 0$$

- Decision function:

$$f_{w,b}(x) = sign\left(\sum_{i=1}^{n} \alpha_i^* y^{(i)} (x^{(i)})^T x + b^*\right)$$

Linearly Separable SVM(Dual)

- **Input:** a linearly separable training set $S = \{(x^{(i)}, y^{(i)}\}_{i=1}^{n}$
- **Output:** a separating hyperplane and decision function
  1. Solving the follwoing optimization problem to obtain the optimal $\alpha^* = (\alpha_1^*, ..., \alpha_n^*)$

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \alpha_i y^{(i)} y^{(j)} (x^{(i)})^T x^{(j)}$$

$$s.t. \quad \alpha_i \geq 0, i = 1, ..., n$$

$$\sum_{i=1}^{n} \alpha_i y^{(i)} = 0$$

  2. Obtain the optimal (w^*, b^*) via the following equations $(\alpha_j^* > 0)$

$$w^* = \sum_{i=1}^{n} \alpha_i^* y^{(i)} x^{(i)}, \quad b^* = y^{(j)} - \sum_{i=1}^{n} \alpha_i^* y^{(i)} \cancel{y^{(j)}} (x^{(i)})^T x^{(j)}$$

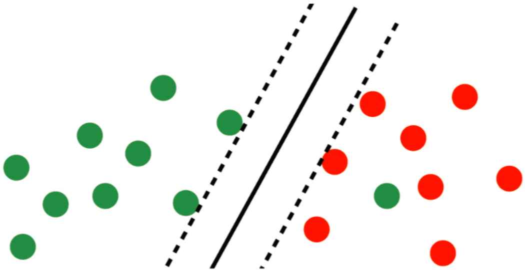
  3. Separating hyperplane: $w^* + b^* = 0$, and the decision function: $f_{w,b}(x) = sign\left(\sum_{i=1}^{n} \alpha_i^* y^{(i)} (x^{(i)})^T x + b^*\right)$

Support Vectors

- Recall that $\alpha_i > 0$ only for support vectors.
- We only need to find the inner products between x and the support vectors in order to calculate w^* and b^* and make our prediction.
- By examining the dual form of the optimization problem, we gained significant insight into the structure of the problem, and were also able to write the entire algorithm in terms of only inner products between input features.
- The property makes it easy to apply the kernel trick to our classification problem, which makes support vector machines efficiently learn in very high dimensional spaces.

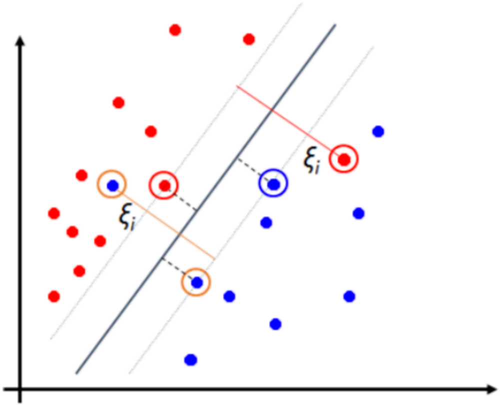
Soft Margin Classification

- What if the training set is not linearly separable?



Soft Margin Classification

- Slack Variables ξ_i can be added to allow misclassification of difficult or noise examples, resulting margin called **soft margin**



Soft Margin Classification

- Introduce a slack variable ξ_i for each example $(x^{(i)}, y^{(i)})$, such that

$$y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i$$

- Adding a penalty: $\sum_{i=1}^{n} \xi_i$
- The optimization problem:

$$\min_{w,b} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} \xi_i$$

$$s.t. \quad y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, \quad i = 1, \ldots, n$$

$$\xi_i \geq 0, \quad i = 1, \ldots, n$$

- Convex quadratic programming problem

Lagrangian Duality for Soft Margin Classification

- Lagrangian:

$$L(w, b, \xi, \alpha, \eta) = \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i[y^{(i)}(w^T x^{(i)} + b) - 1 + \xi_i] - \sum_{i=1}^{n}\eta_i\xi_i$$

- The α_is and η_is are the Lagrangian Multipliers
- Minimize $L(w, b, \xi, \alpha, \eta)$ with respect to w, b and ξ (for fixed α, η) to get $\theta_D(\alpha, \eta)$
- Find d^* by $\max_{\alpha,\eta} \theta_D(\alpha, \eta)$

Lagrangian Duality for Soft Margin Classification

- Minimize $L(w, b, \xi, \alpha, \eta)$ with respect to w, b and ξ to get $\theta_D(\alpha, \eta)$
  - Setting the derivatives of L with respect to w, b and ξ to zero:

$$\nabla_w L(w, b, \xi, \alpha, \eta) = w - \sum_{i=1}^{n} \alpha_i y^{(i)} x^{(i)} = 0$$

  - For the derivative with respect to b, we can obtain:

$$\frac{\partial L(w, b, \xi, \alpha, \eta)}{\partial b} = \sum_{i=1}^{n} \alpha_i y^{(i)} = 0$$

  - For the derivative with respect to ξ, we can obtain:

$$\frac{\partial L(w, b, \xi, \alpha, \eta)}{\partial \xi_i} = C - \alpha_i - \eta_i = 0$$

Lagrangian Duality for Soft Margin Classification

- Plug them back into the lagrangian, we get:

$$\theta_D(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y^{(i)} y^{(j)} \alpha_i \alpha_j (x^{(i)})^T x^{(j)}$$

- Dual form of the problem:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y^{(i)} y^{(j)} \alpha_i \alpha_j (x^{(i)})^T x^{(j)}$$

$$s.t. \quad 0 \leq \alpha_i \leq C, \quad i = 1, ..., n$$

$$\sum_{i=1}^{n} = 0$$

Soft Margin Classifier

Theorem (Solution of the Primal Optimization Problem)

Suppose that $\alpha^* = (\alpha_1^*, ..., \alpha_n^*)$ *are the optimal solution of the dual optimization problem, then there exists* j *such that* $0 < \alpha_j^* < C$, *and we have,*

$$w^* = \sum_{i=1}^{n} \alpha_i^* y^{(i)} x^{(i)}, \quad b^* = y^{(j)} - \sum_{i=1}^{n} \alpha_i^* y^{(i)} \cancel{y^{(j)}} (x^{(i)})^T x^{(j)}$$

- Separating hyperplane:

$$\sum_{i=1}^{n} \alpha_i^* y^{(i)} (x^{(i)})^T x + b^* = 0$$

- Decision function:

$$f_{w,b}(x) = sign\left(\sum_{i=1}^{n} \alpha_i^* y^{(i)} (x^{(i)})^T x + b^*\right)$$

Non-Separable SVM(Dual)

- **Input:** a linearly separable training set $S = \{(x^{(i)}, y^{(i)}\}_{i=1}^{n}$
- **Output:** a separating hyperplane and decision function
  1. Choose parameter C and solving the following optimization problem to obtain the optimal $\alpha^* = (\alpha_1^*, ..., \alpha_n^*)$

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \alpha_i y^{(i)} y^{(j)} (x^{(i)})^T x^{(j)}$$

$$s.t. \;\; 0 < \alpha_i \leq C, \;\; i = 1, ..., n$$

$$\sum_{i=1}^{n} \alpha_i y^{(i)} = 0$$

  2. Obtain the optimal (w^*, b^*) via the following equations $(0 < \alpha_j^* < C)$

$$w^* = \sum_{i=1}^{n} \alpha_i^* y^{(i)} x^{(i)}, \quad b^* = y^{(j)} - \sum_{i=1}^{n} \alpha_i^* y^{(i)} \cancel{y^{(j)}} (x^{(i)})^T x^{(j)}$$

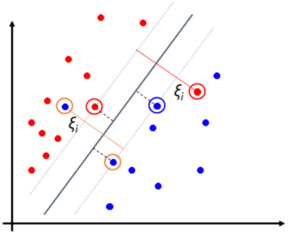
  3. Separating hyperplane: $w^* + b^* = 0$, and the decision function: $f_{w,b}(x) = sign\left(\sum_{i=1}^{n} \alpha_i^* y^{(i)} (x^{(i)})^T x + b^*\right)$

Support Vectors in Non-separable Case

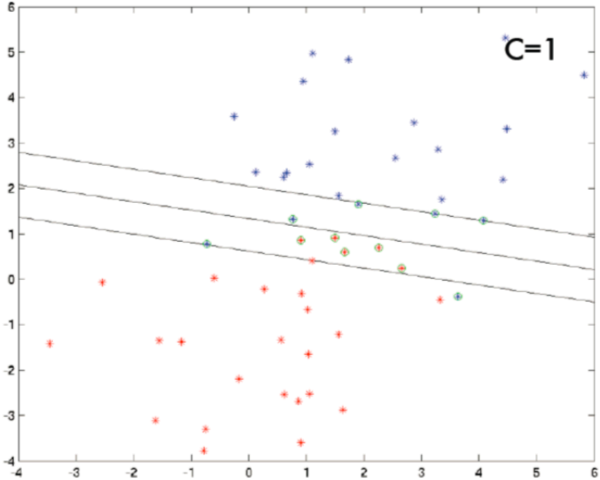
- We call the training examples $(x^{(i)}, y^{(i)})$ as the support vectors, if the corresponding $\alpha_i^* > 0$.
- Recall the KKT dual complementarity condition $\alpha_i * g_i(w^*) = 0$, and $\eta_i * \xi_i = 0$. That is:

$$\alpha_i^*[y^{(i)}(w^{*T}x^{(i)} + b) - 1 + \xi_i] = 0, \quad \eta_i * \xi_i = 0.$$

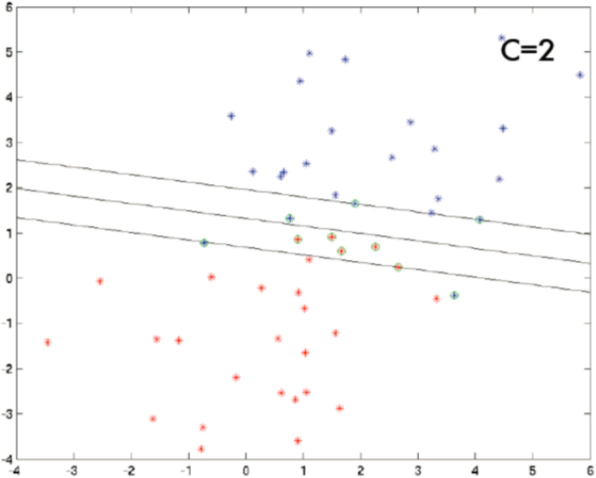
- If $\alpha_i^* = 0$, we have $y^{(i)}(w^{*T}x^{(i)} + b) \geq 1$. (in the correct sides)
- If $0 < \alpha_i^* < C$, we have $y^{(i)}(w^{*T}x^{(i)} + b) = 1$. (ontheboundary)
- If $\alpha_i^* = C$, we have $y^{(i)}(w^{*T}x^{(i)} + b) \leq 1$. (in the wrong sides)



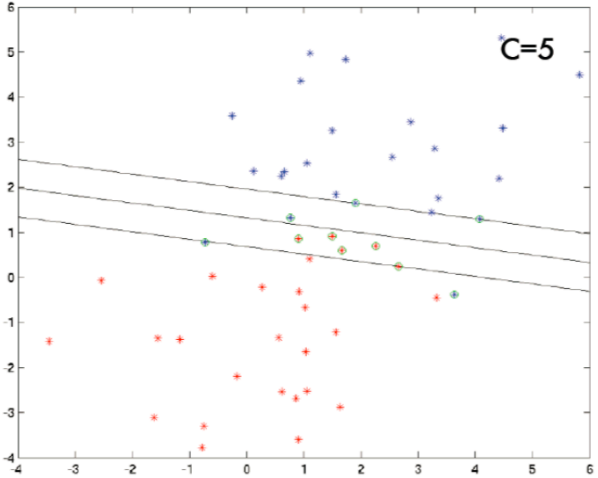
Parameter C



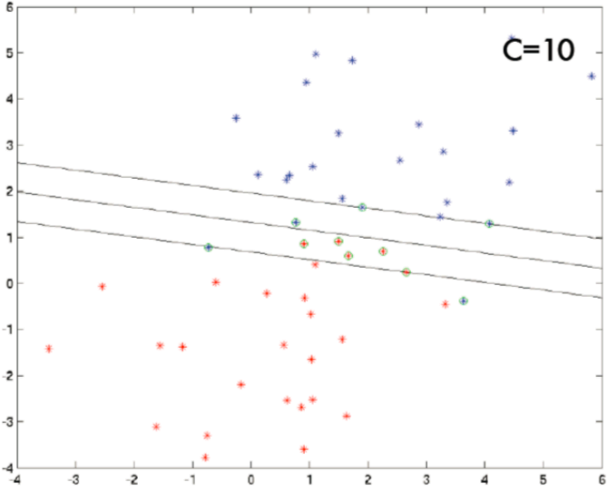
Parameter C



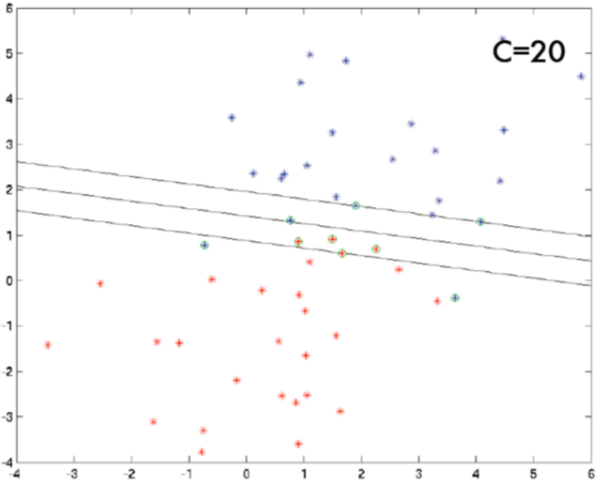
Parameter C



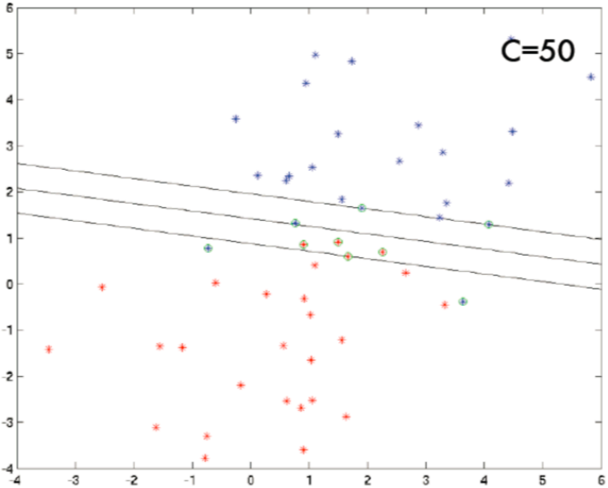
Parameter C



Parameter C



Parameter C



Parameter C