Support Vector Machine I

Yanyan Lan

Institute of Computing Technology, Chinese Academy of Sciences

lanyanyan@ict.ac.cn

November 16, 2015

Questions and Answers 1

- Naive Bayes
  - Question: $P(y=1|x=(2,S)')+P(y=-1|x=(2,S)')\neq 1$?
  - Answer: $P(y=1|x=(2,S)')+P(y=-1|x=(2,S)')=1$.
  - Reason:

$$P(y=1|x=(2,S)')=\frac{P(x=(2,S)',y=1)}{P(x=(2,S)',y=1)+P(x=(2,S)',y=-1)},$$

$$P(y=-1|x=(2,S)')=\frac{P(x=(2,S)',y=-1)}{P(x=(2,S)',y=1)+P(x=(2,S)',y=-1)}$$

  - Independent Assumption: $P(x|y)=P(x_1,x_2|y)=P(x_1|y)P(x_2|y)$.
  - Another Way: $\max\{P(x=2,S,y=1),P(x=2,S,y=1)\}$ $\max\{P(y=1)P(x_1=2|y=1)P(x_2=S|y=1),P(y=-1)P(x_1=2|y=-1)P(x_2=S|y=-1)\}$.
  - Note: $P(x=(2,S))=P(x_1=2)P(x_2=S)$?

Questions and Answers 2

- Learning Rate Adaptation in SGD

$$\omega^{t+1} = \omega^{t} - \alpha_t \nabla Q(\omega^t, x, y).$$

  - Bold Driver: After each epoch, compare the loss to int previous value. If the error has decreased, increase α by a small proportion (1% to 5%). If the error has increased by more than a tiny proportion (e.g.10^{-6}), decrease α sharply (typically 50%).
  - Evaluate a hold out set after each epoch and anneal the learning rate when the change in objective between epochs is below a small threshold.
  - Annealing:

$$\alpha_t = \frac{\alpha_0}{1 + t/T},$$

    it will keeps α_t nearly constant for the first T training epochs, then anneal it at a very slow pace that is know from theory to guarantee convergence to the minimum. T is a new free parameter to be determined by trial and error.

Question and Answers 3

- Bias and Variance Decomposition

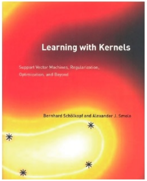
$$
\begin{aligned}
E(f_D) = E_D \int (f_D(x) - y)^2 p(x,y) dxdy \\
= \int (E_D(f_D(x)) - h(x))^2 p(x) dx \ (\textit{Bias}^2) \\
+ \int E_D(f_D(x) - E_D(f_D(x)))^2 p(x) dx \ (\textit{Variance}) \\
+ \int \int (h(x) - y)^2 p(x,y) dxdy \ (\textit{Noise}).
\end{aligned}
$$

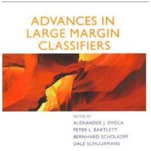
Support Vector Machine: History

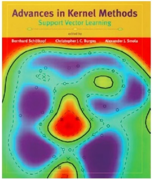
- Linear SVM: Cortes and Vapnik, Support Vector Networks, *Machine Learning*, 1995.
- Kernelized SVM: Boster, Guyon, and Vapnik, A Training Algorithm for Optimal Margin Classifiers, *workshop in COLT*, 1992.
- SVR: Drucker, burges, Smola, and Vapnik, Support Vector Regression Machine, *NIPS* 1996.
- 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

Support Vector Machine:References

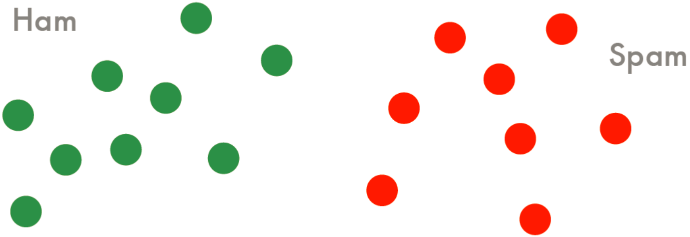
- Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond,
- Advances in Large Margin Classifiers,
- Advances in Kernel Methods: Support Vector Learning.


Learning with Kernels
Support Vector Machines, Regularization, Optimization, and Beyond
Bernhard Schölkopf and Alexander J. Smola

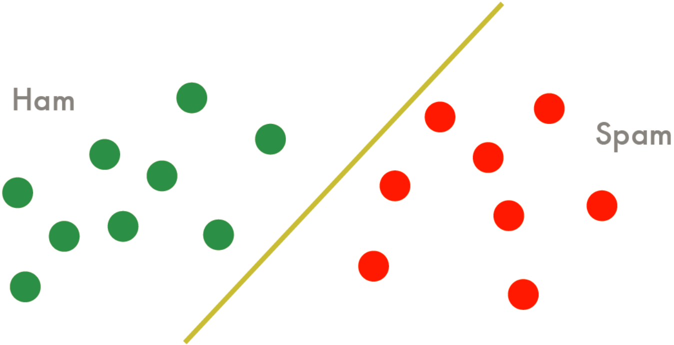




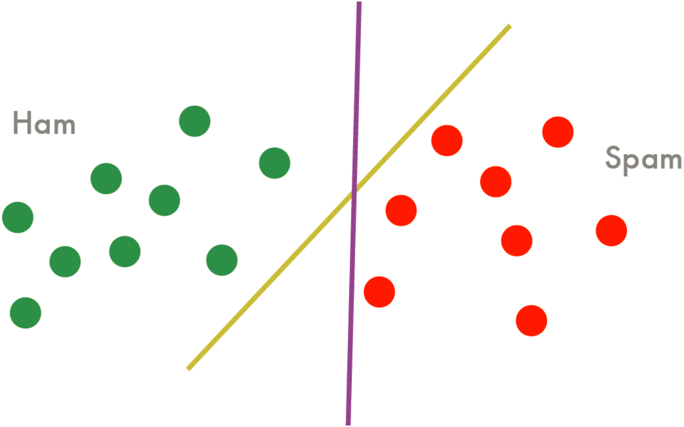
Linearly Separable Classification



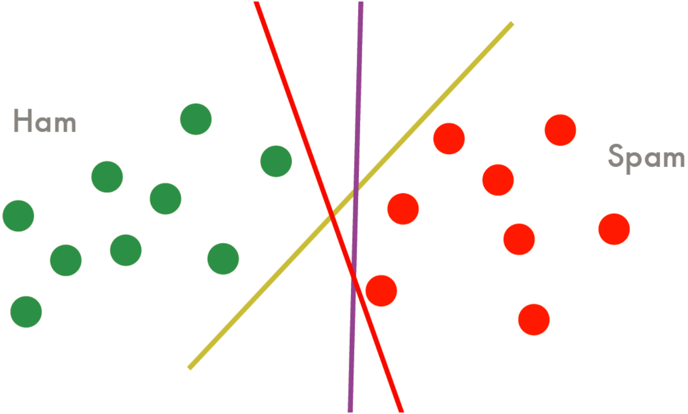
Linearly Separable Classification



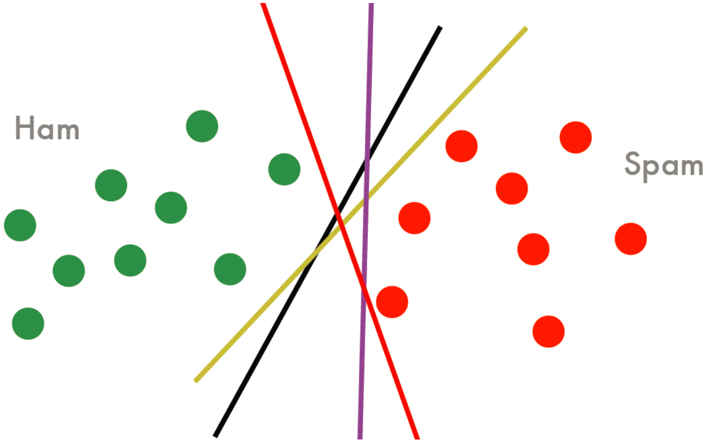
Linearly Separable Classification



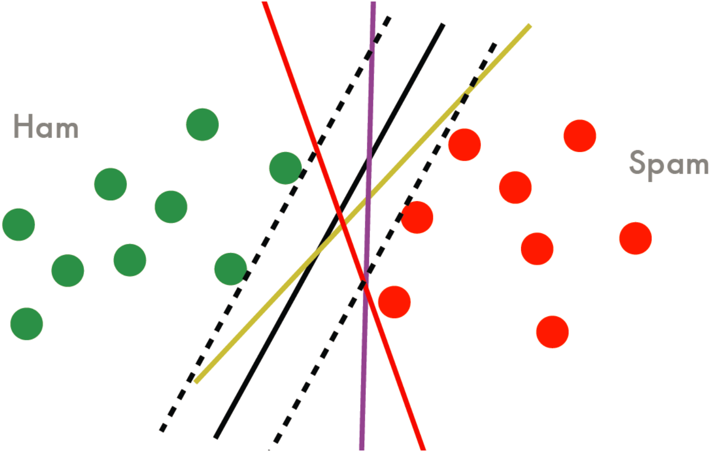
Linearly Separable Classification



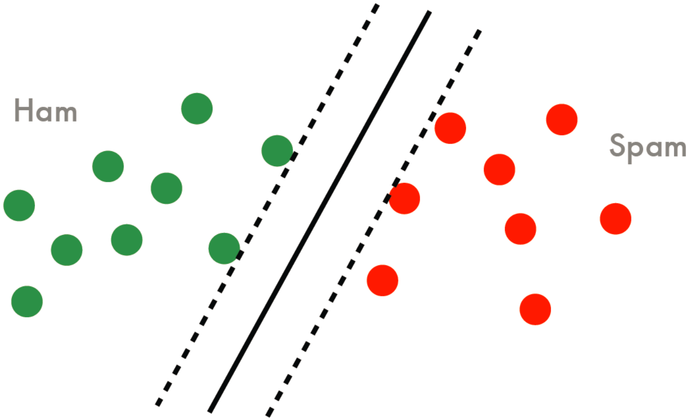
Linearly Separable Classification



Linearly Separable Classification



Optimal Classifier



Margin: Intuition

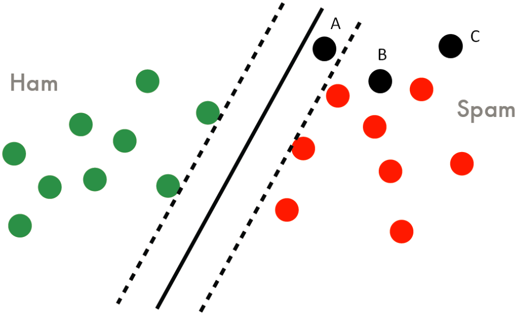
- Logistic Regression

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

  - Predict 1 if $P(y = 1|x) \geq 0.5$, i.e. $\omega^T x \geq 0$;
  - Predict 0 if $P(y = 1|x) < 0.5$, i.e. $\omega^T x < 0$;
  - The larger $\omega^T x$ 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 ω so that $\omega^T x >> 0$ whenever $y = 1$, and $\omega^T x << 0$ 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:

$$f_{\omega,b}(x) = g(\omega^T x + b).$$

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

Functional Margin

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

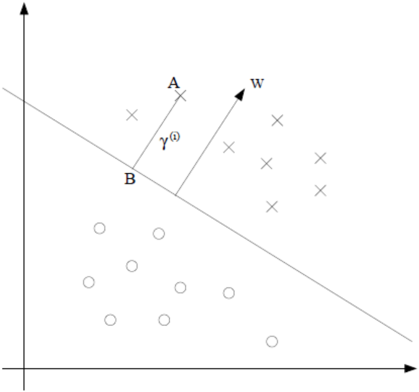
$$\hat{\gamma}^{(i)} = y^{(i)}(\omega^T x^{(i)} + b)$$

  - If $y^{(i)} = 1$, we need $\omega^T x^{(i)} + b$ to be a large positive number.
  - If $y^{(i)} = -1$, we need $\omega^T x^{(i)} + b$ to be a large negative number.
  - If $y^{(i)}(\omega^T x^{(i)} + b) > 0$, then our prediction on this example is correct.
  - A large functional margin represents a confident and a correct prediction.

- Given a training set $S = \{(x^{(i)}, y^{(i)}), i = 1, \cdots, n\}$, the functional margin of (ω, 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 $x^{(i)} - \gamma^{(i)}\omega/\|\omega\|_2$.
- B lies in the decision boundary:

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

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

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

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

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

Geometric Margin

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

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

- The relationships between functional margin and geometric margin:

$$\gamma^{(i)} = \frac{\hat{\gamma^{(i)}}}{\|\omega\|_2}, \ \gamma = \frac{\hat{\gamma}}{\|\omega\|_2}.$$

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

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

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

- 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.)

$$\max_{\hat{\gamma},\omega,b} \quad \frac{\hat{\gamma}}{\|\omega\|_2}$$
$$s.t. \quad y^{(i)}(\omega^T x^{(i)} + b) \geq \hat{\gamma}, i = 1, \cdots, n.$$

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

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

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

Linearly Separable SVM

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

$$\begin{aligned} \min_{\omega, b} \quad & \frac{1}{2}\|\omega\|_2^2 \\ s.t. \quad & y^{(i)}(\omega^T x^{(i)} + b) \geq 1, i = 1, \cdots, n, \end{aligned}$$

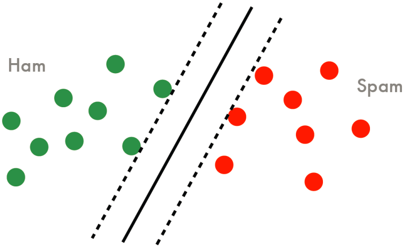
  2. Separating hyperplane: $\omega^* x + b^* = 0$, and decision function: $f_{\omega,b}(x) = sign({\omega^*}^T x + b^*)$.

Theoretical Guarantee

The existence and uniqueness of max margin classifier both hold for linearly separable training set.

Support Vectors and Margin

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



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

Lagrange Duality: Simplest Case

- Equality constrained optimization problem:

$$\begin{aligned} \min_{\omega} \quad & f(\omega) \\ s.t. \quad & h_i(\omega) = 0, i = 1, \cdots, l, \end{aligned}$$

- Lagrangian:

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

  - The β_is are called the Lagrange Multipliers.

- We would then find and set L's partial derivatives to zero:

$$\frac{\partial L}{\partial \omega_i} = 0, \quad \frac{\partial L}{\partial \beta_i} = 0.$$

Lagrange Duality: Generalized Case

- Optimization problem with both inequality and equality constraints:

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

  - Primal optimization problem.
- Generalized Lagrangian:

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

  - The αs and βs are the Lagrange Multipliers.

Lagrange Duality

- Consider the quantity:

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

  - If ω violates any of the primal constraints, then:

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

  - If the constraints are indeed satisfied, then:

$$\theta_P(\omega) = f(\omega).$$

Lagrange Duality

- Consider the minimization problem:

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

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

Lagrange Duality

- Consider a different problem:

$$\theta_D(\alpha, \beta) = \min_{\omega} L(\omega, \alpha, \beta).$$

- Dual optimization problem:

$$\max_{\alpha,\beta:\alpha_i \geq 0} \theta_D(\alpha, \beta) = \max_{\alpha,\beta:\alpha_i \geq 0} \min_{\omega} L(\omega, \alpha, \beta).$$

- 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:

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

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

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

$$\begin{aligned}
\frac{\partial}{\partial \omega_i} L(\omega^*, \alpha^*, \beta^*) &= 0,\ i = 1, \cdots, M, \\
\frac{\partial}{\partial \beta_i} L(\omega^*, \alpha^*, \beta^*) &= 0,\ i = 1, \cdots, l, \\
\alpha_i^* g_i(\omega^*) &= 0,\ i = 1, \cdots, k, \ \textit{KKT dual complementarity} \\
g_i(\omega^*) &\leq 0,\ i = 1, \cdots, k, \\
\alpha_i^* &\geq 0,\ i = 1, \cdots, k.
\end{aligned}$$

- If some $\omega^*, \alpha^*, \beta^*$ satisfy the KKT condition, then it is also a solution to the primal and dual problems.

Optimal Margin Classifier: Dual Solution

- Primal optimization problem for finding the optimal margin classifier:

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

- The constraint can be written as:

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

- 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.,$g_i(\omega) = 0$). 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:

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

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

Optimal Margin Classifier: Dual Solution

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

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

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

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

  - Plug them back into the lagrangian, we get:

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

Optimal Margin Classifier: Dual Solution

- Dual optimization problem:

$$\begin{aligned}
\max_{\alpha} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y^{(i)} y^{(j)} \alpha_i \alpha_j (x^{(i)})^T x^{(j)} \\
s.t. \quad & \alpha_i \geq 0, i = 1, \cdots, n, \\
& \sum_{i=1}^{n} \alpha_i y^{(i)} = 0.
\end{aligned}$$

Optimal Margin Classifier: Primal Solution via Dual Optimization

Theorem: Solution of the Primal Optimization Problem

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

$$\omega^* = \sum_{i=1}^{n} \alpha_i^* y^{(i)} x^{(i)}, \; b^* = y^{(j)} - \sum_{i=1}^{n} \alpha_i^* y^{(i)} (x^{(i)})^T x^{(j)}.$$

- Separating hyperplane:

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

- Decision function:

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

Linearly Separable SVM (Dual)

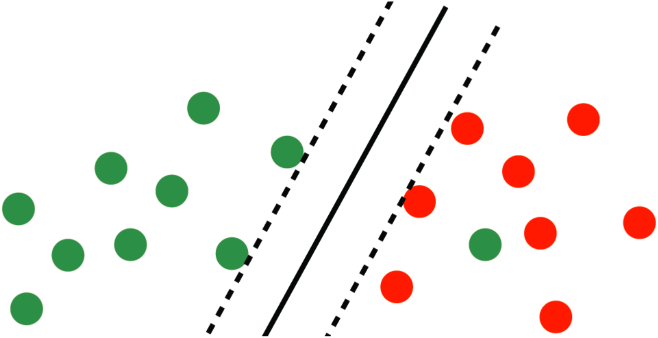
- Input: a linearly separable training set $S = \{(x^{(i)}, y^{(i)}), i = 1, \cdots, n\}$;
- Output: a separating hyperplane and decision function.
  1. Solving the following optimization problem to obtain the optimal $\alpha^* = (\alpha_1^*, \cdots, \alpha_l^*)$:
  $$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y^{(i)} y^{(j)} \alpha_i \alpha_j (x^{(i)})^T x^{(j)}$$
  $$s.t. \quad \alpha_i \geq 0, i = 1, \cdots, n,$$
  $$\sum_{i=1}^{n} \alpha_i y^{(i)} = 0.$$
  2. Obtain the optimal (ω^*, b^*) via the following equations $(\alpha_j^* > 0)$:
  $$\omega^* = \sum_{i=1}^{n} \alpha_i^* y^{(i)} x^{(i)}, \ b^* = y^{(j)} - \sum_{i=1}^{n} \alpha_i^* y^{(i)} (x^{(i)})^T x^{(j)}.$$
  3. Separating hyperplane: $\omega^* x + b^* = 0$, and decision function: $f_{\omega,b}(x) = sign({\omega^*}^T x + b^*)$.

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.

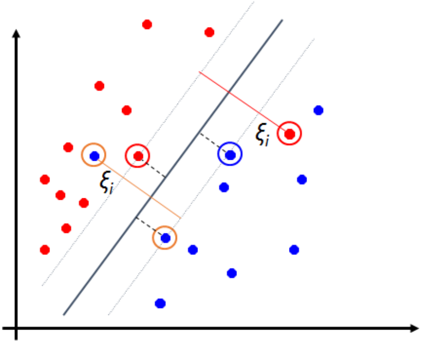
Soft Margin Classification

- 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 noisy examples, resulting margin called soft margin.



Soft Margin Classification

- Introduce a slack variable $\xi_i \geq 0$ for each example $(x^{(i)}, y^{(i)})$, such that:

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

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

$$\begin{aligned} \min_{\omega, b} \quad & \frac{1}{2}\|\omega\|_2^2 + C\sum_{i=1}^{n} \xi_i \\ s.t. \quad & y^{(i)}(\omega^T x^{(i)} + b) \geq 1 - \xi_i, \ \ i = 1, \cdots, n, \\ & \xi_i \geq 0, \ \ i = 1, \cdots, n. \end{aligned}$$

- Convex quadratic programming problem.

Lagrangian Duality for Soft Margin Classification

- Lagrangian:

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

  - The α_is and η_is are our Lagrangian Multipliers.

1. Minimize $L(\omega, b, \xi, \alpha, \eta)$ with respect to ω, b and ξ(for fixed α, η) to get $\theta_D(\alpha, \eta)$.
2. Find d^* by $\max_{\alpha,\eta}\theta_D(\alpha, \eta)$.

Lagrangian Duality for Soft Margin Classification

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

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

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

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

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

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

  - Plug them back into the lagrangian, we get:

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

Lagrangian Duality for Soft Margin Classification

- Dual form of the problem:

$$
\begin{aligned}
\max_{\alpha} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y^{(i)} y^{(j)} \alpha_i \alpha_j (x^{(i)})^T x^{(j)} \\
s.t. \quad & 0 \le \alpha_i \le C, i = 1, \cdots, n, \\
& \sum_{i=1}^{n} \alpha_i y^{(i)} = 0.
\end{aligned}
$$

Soft Margin Classifier

Solution of the Primal Optimization Problem

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

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

- Separating hyperplane:

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

- Decision function:

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

Non-Separable SVM (Dual)

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

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

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

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

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

$$\omega^* = \sum_{i=1}^{n} \alpha_i^* y^{(i)} x^{(i)}, \;\; b^* = y^{(j)} - \sum_{i=1}^{n} \alpha_i^* y^{(i)} (x^{(i)})^T x^{(j)}.$$

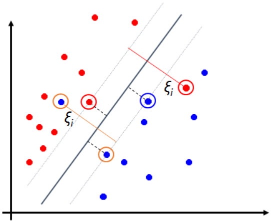
  3. Separating hyperplane: $\omega^* x + b^* = 0$, and decision function: $f_{\omega,b}(x) = sign({\omega^*}^T x + b^*)$.

Support Vectors in Non-Separable Case

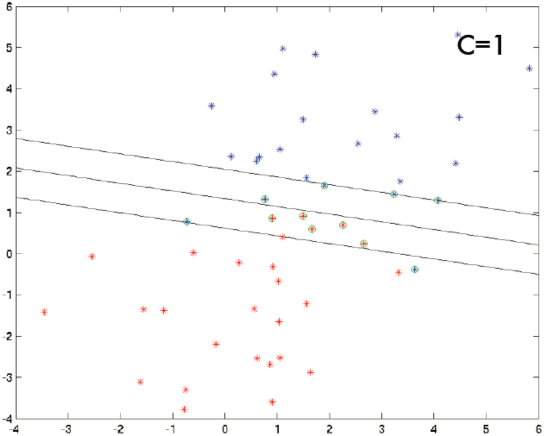
- We call the training examples $(x^{(i)}, y^{(i)})$ as the support vectors, if the corresponding $\alpha_i^* > 0$.
- Recall the KKT dual complementarity condition $\alpha_i^* g_i(\omega^*) = 0$, and $\eta_i^* \xi_i = 0$. That is:

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

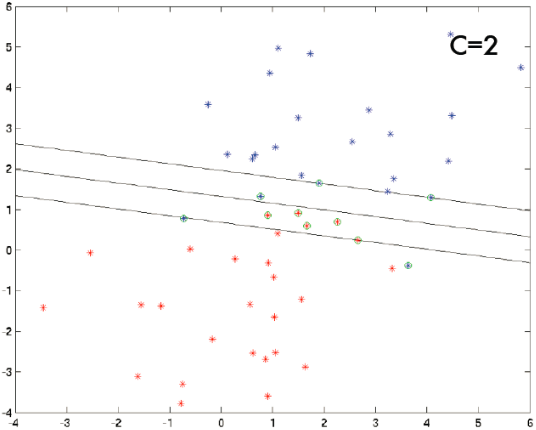
1. If $\alpha_i^* = 0$, we have $y^{(i)}({w^*}^T x^{(i)} + b) \geq 1$. (in the correct sides)
2. If $0 < \alpha_i^* < C$, we have $y^{(i)}({w^*}^T x^{(i)} + b) = 1$. (on the boundary)
3. If $\alpha_i^* = C$, we have $y^{(i)}({w^*}^T x^{(i)} + b) \leq 1$. (in the wrong sides)



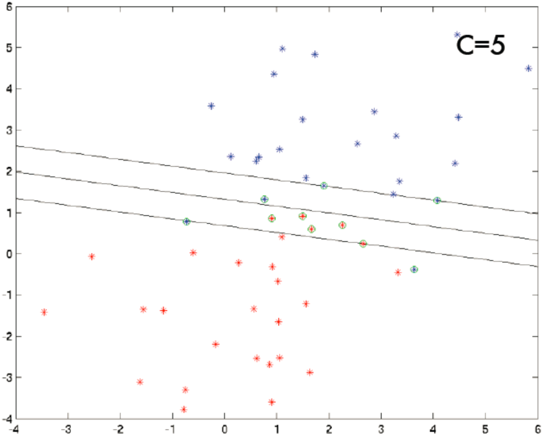
Parameter C



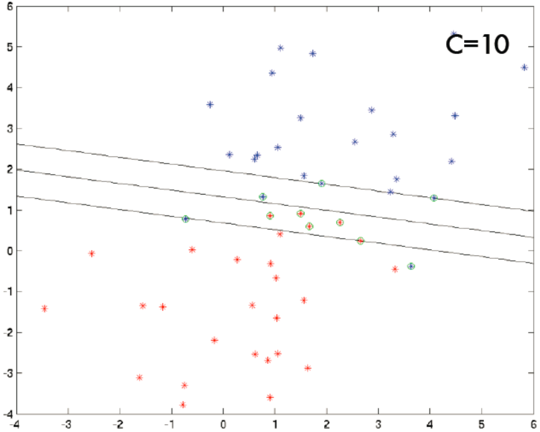
Parameter C



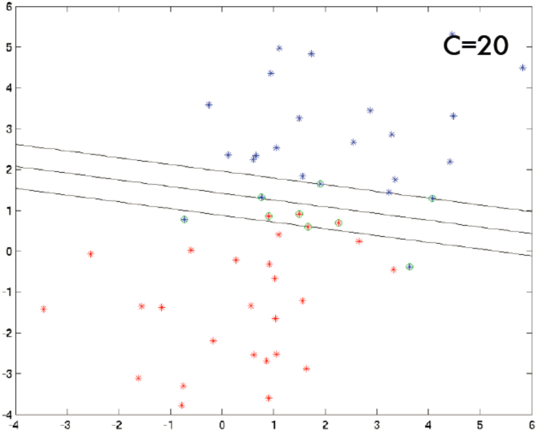
Parameter C



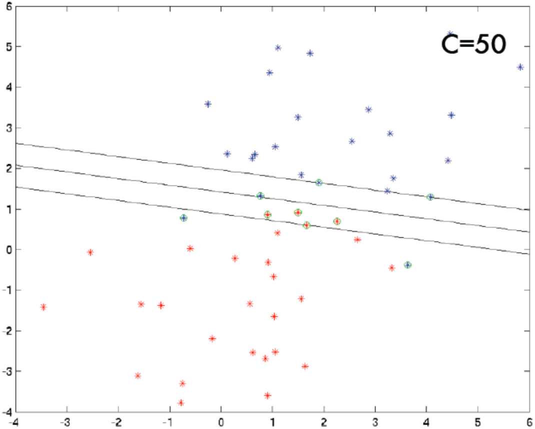
Parameter C



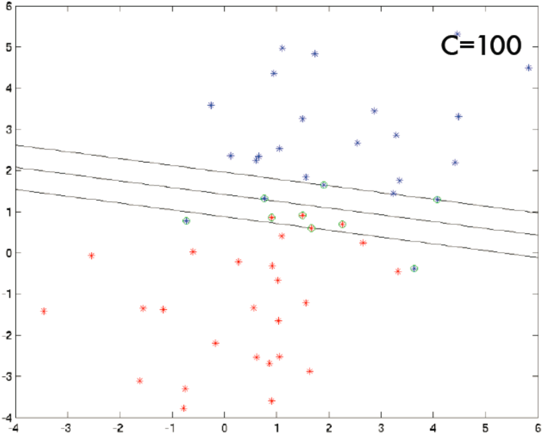
Parameter C



Parameter C



Parameter C



Loss Function of SVM

- Optimization Aspect:

$$\begin{aligned} \min_{\omega,b} \quad & \frac{1}{2}\|\omega\|_2^2 + C\sum_{i=1}^{n}\xi_i \\ s.t. \quad & y^{(i)}(\omega^T x^{(i)} + b) \geq 1 - \xi_i,\ \ i = 1, \cdots, n, \\ & \xi_i \geq 0,\ \ i = 1, \cdots, n. \end{aligned}$$

- Loss Function Aspect:

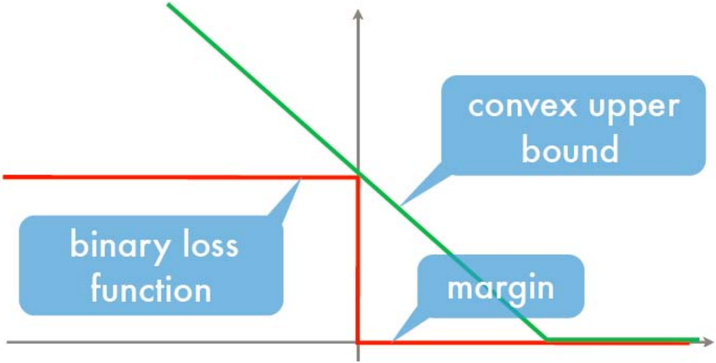
$$\min_{\omega,b} \sum_{i=1}^{n}[1 - y^{(i)}(\omega^T x^{(i)} + b)]_+ + \lambda\|\omega\|_2^2.$$

- We can prove that the above two optimization problem are equivalent.

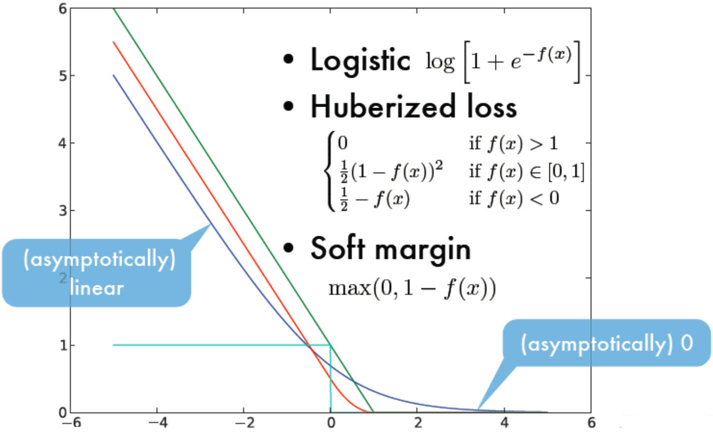
Hinge Loss as a Surrogate of the Binary Loss

- The loss function of SVM is:

$$L(f, x, y) = [1 - y(\omega^T x + b)]_+ .$$ Hinge Loss Function



Loss Comparison



- **Logistic** $\log\left[1+e^{-f(x)}\right]$
- **Huberized loss**

$$\begin{cases} 0 & \text{if } f(x) > 1 \\ \frac{1}{2}(1-f(x))^2 & \text{if } f(x) \in [0,1] \\ \frac{1}{2} - f(x) & \text{if } f(x) < 0 \end{cases}$$

- **Soft margin**

$$\max(0, 1 - f(x))$$

(asymptotically) linear

(asymptotically) 0

- Please note that $f(x)$ in the figure stands for $yf(x)$ as in our notation.

Question and Answering