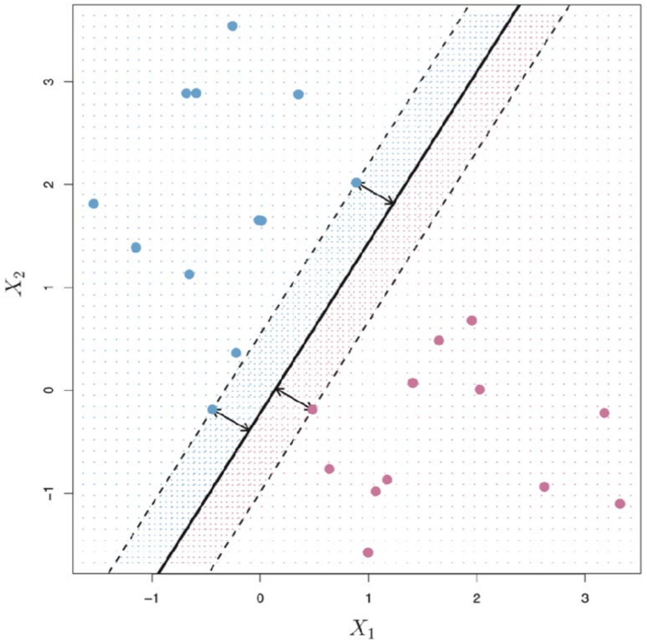
CSCE 633: Machine Learning

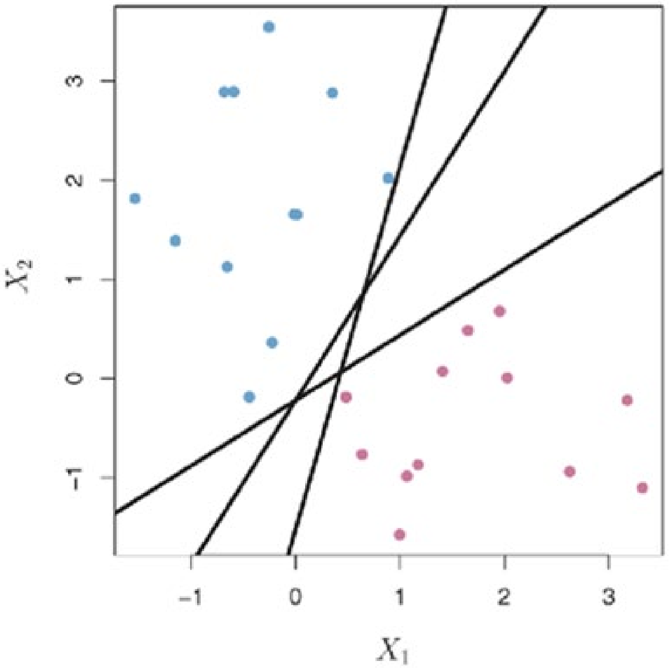
Lecture 22: SVM Optimization

Texas A&M University
Bobak Mortazavi

Review

1. Which of the decision boundaries would you use from the plot on the right and why?
2. In the plot on the left, what are the dashed lines called?
3. In the plot of the left what are the dots on the dashed line called?
4. How can we modify the Maximum Margin Classifier for data that isn't linearly separable?





Understanding the Loss Function

- Consider the margin boundary $y_i f(x) = y_i(w_0 + \sum_{j=1}^{D} w_j\, \phi(x_{ij}))$

Understanding the Loss Function

- Consider the margin boundary $y_i f(x) = y_i(w_0 + \sum_{j=1}^{D} w_j\, \phi(x_{ij}))$
- Why do we multiply y_i by $f(x)$?

Understanding the Loss Function

- Consider the margin boundary $y_i f(x) = y_i(w_0 + \sum_{j=1}^{D} w_j\, \phi(x_{ij}))$
- Why do we multiply y_i by $f(x)$?
- Can we somehow relate SVM's margin boundary to Regression's Loss Functions?

Understanding the Loss Function

- Consider the margin boundary $y_i f(x) = y_i(w_0 + \sum_{j=1}^{D} w_j \, \phi(x_{ij}))$
- Why do we multiply y_i by $f(x)$?
- Can we somehow relate SVM's margin boundary to Regression's Loss Functions?
- Recall $y - \hat{y} = y - f(x)$ is our residual (error)

Classification Rule

- The classification rule for SVM is $G(x) = \text{sign}(f(x))$

Classification Rule

- The classification rule for SVM is $G(x) = \text{sign}(f(x))$
- Now, how do we classify error?
- Recall $y_i \in \{-1, +1\}$

Classification Rule

- The classification rule for SVM is $G(x) = \text{sign}(f(x))$
- Now, how do we classify error?

Classification Rule

- The classification rule for SVM is $G(x) = \text{sign}(f(x))$
- Now, how do we classify error?
- Recall $y_i \in \{-1, +1\}$
- So, $y_i G(x_i) > 0$ if samples are classified correctly

0-1 Loss

- The decision boundary as $f(x) = 0$

0-1 Loss

- The decision boundary as $f(x) = 0$
- $L(y, f(x))$ is called the 0-1 loss in this case

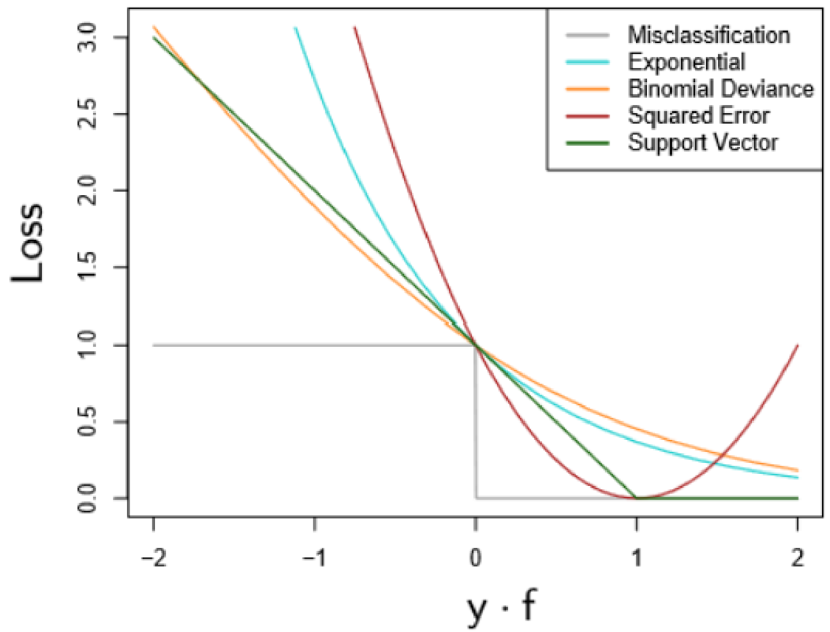
0-1 Loss

- The decision boundary as $f(x) = 0$
- $L(y, f(x))$ is called the 0-1 loss in this case
- $L\big(y, f(x)\big) = \sum_{i=1}^{N} I(y_i f(x_i) < 0)$

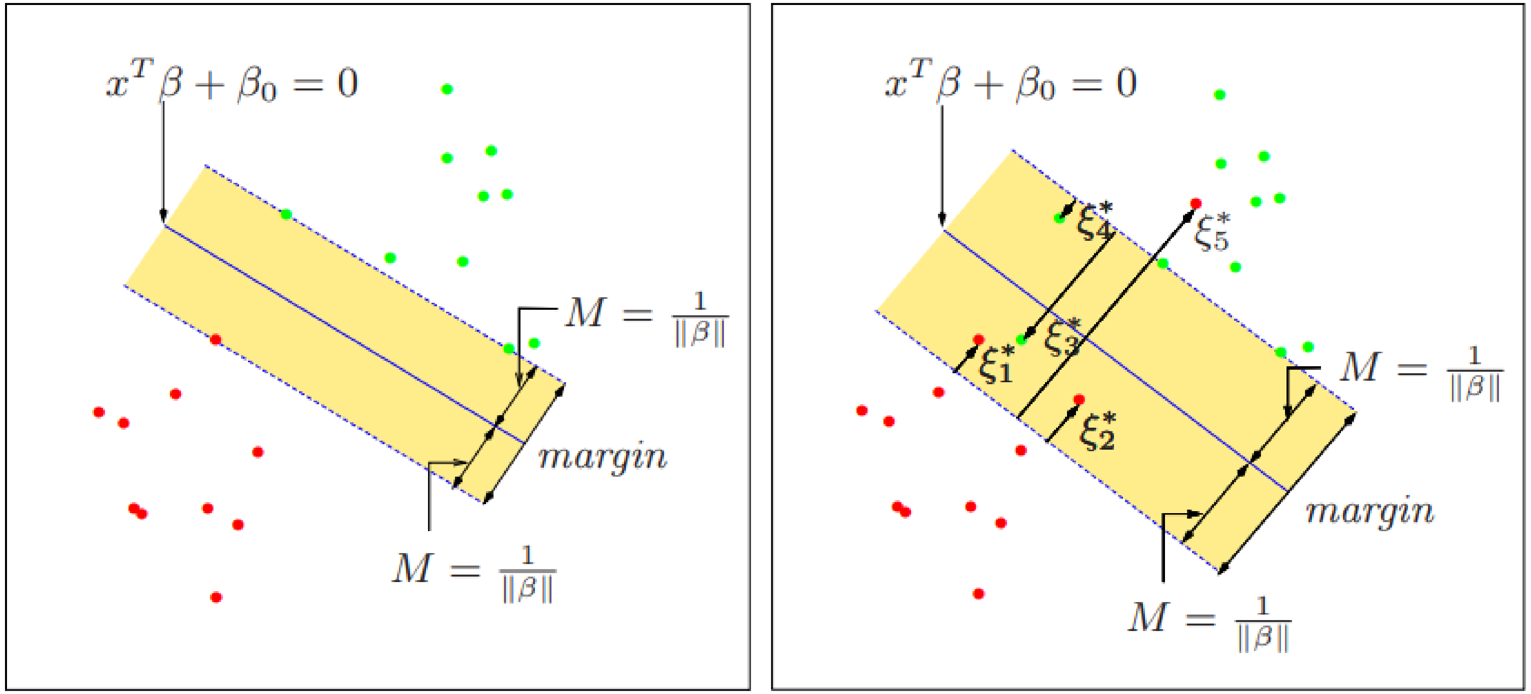
Support Vector Machine: Hinge Loss

$$L(x, y, w) = \sum_{i=1}^{N} \max(0, 1 - y_i(w_0 + w_1 x_{i1} + \cdots + w_D x_{iD}))$$

- Instead of the common loss for logistic regression



Optimal Hyperplane and Support Vectors



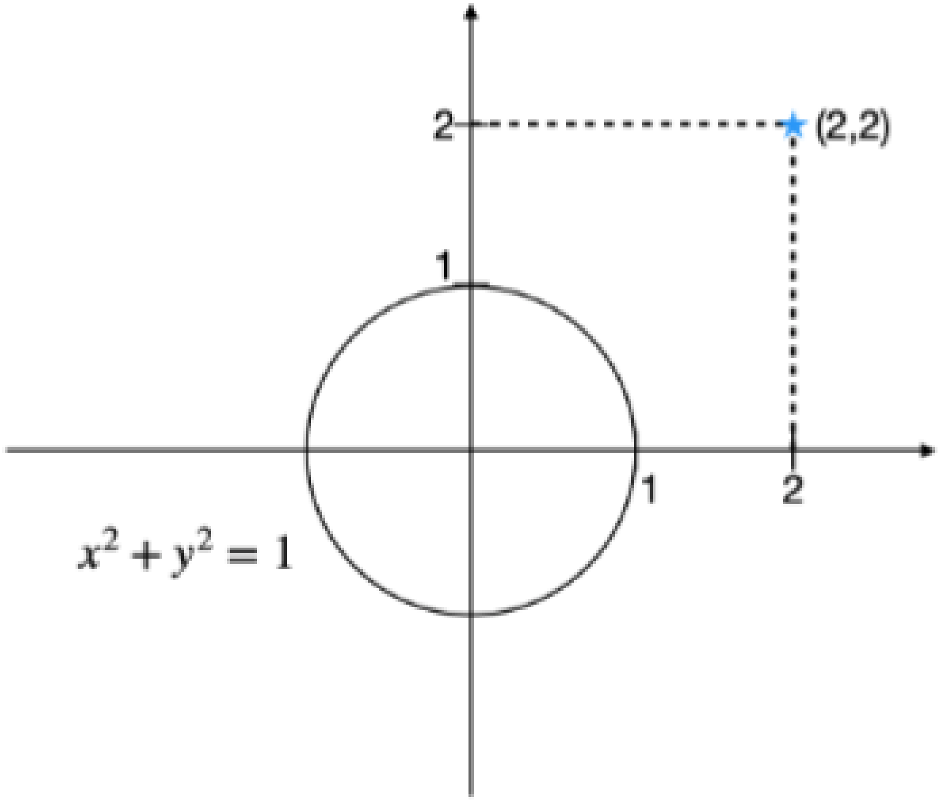
- Margin of Separation M: distance between the separating hyperplane and the closest input point
- Support Vectors: input points closest to the separating hyperplane

Mathematical Aside: Lagrange Multipliers

- Turn a constrained optimization problem into an unconstrained optimization problem by absorbing the constraints into a cost function, weighted by the Lagrange multipliers
- Example: Find point on the circle $x^2 + y^2 = 1$ closest to the point (2,2)
  – Minimize $F(x, y) = (x-2)^2 + (y-2)^2$
  – Subject to the constraint $x^2 + y^2 - 1 = 0$
  – Absorb the constraint into the cost function, after multiplying the Lagrange multiplier α:

$$F(x, y, \alpha) = (x-2)^2 + (y-2)^2 + \alpha(x^2 + y^2 - 1)$$

Mathematical Aside: Langrange Multipliers



Mathematical Aside: Lagrange Multipliers

- Formulate Lagrangian (primal problem):
- $F(x, y, \alpha) = (x-2)^2 + (y-2)^2 + \alpha(x^2 + y^2 - 1)$
- The optimization problem becomes:

$$\frac{\partial F}{\partial x} = 2(x-2) + 2\alpha \mathrm{x} = 0 \rightarrow x = \frac{2}{1+\alpha}$$

$$\frac{\partial F}{\partial y} = 2(y-2) + 2\alpha \mathrm{y} = 0 \rightarrow y = \frac{2}{1+\alpha}$$

- We substitute x,y in the Lagrangian and express it in terms of its dual form wrt α and maximize it

$$\frac{\partial F}{\partial x} = x^2 + y^2 - 1 = 0 \rightarrow \left(\frac{2}{1+\alpha}\right)^2 + \left(\frac{2}{1+\alpha}\right)^2 = 1 \rightarrow \alpha = 2\sqrt{2} - 1$$

- Recover the solution: $(x, y) = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$

Mathematical Aside: Lagrange Multipliers

- Exercise
  - Find point on the circle $x^2 + y^2 = 1$ closest to the point (-3,3)

Primal Problem: Constrained Optimization

- For the training set $D^{train} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ find w and w_0 such that they minimize the inverse separation margin $\left(\frac{1}{M} = \frac{\|w\|}{2}\right)$ while satisfying a constraint (all examples are correctly classified):
  - Cost function $\Phi(w) = \frac{1}{2} w^T w$
  - Constraint: $y_i(w^T x_i + w_0) \geq 1$ for $i = 1,2,\cdots,N$

$$\min_{w} \frac{1}{2} w^T w, \text{ such that (s.t.) } y_i(w^T x_i + w_0) \geq 1 \text{ for } i = 1,2,\cdots,N$$

- This problem can be solved using the method of Lagrange multipliers (see next two slides)

Support Vector Machines: Linearly separable case

$$\min_{w} \frac{1}{2} w^T w, \text{ such that (s.t.) } y_i(w^T x_i + w_0) \geq 1 \text{ for } i = 1,2,\cdots,N$$

1. Formulate Lagrangian function (primal problem)

$$L = \frac{1}{2}\|\boldsymbol{w}\|_2^2 - \sum_{i=1}^{N} \alpha_i \left(y_i (\boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{x}_i + w_0) - 1 \right)$$

2. Minimize Lagrangian to solve for primal variables w and w_0

$$\frac{\partial L}{\partial \boldsymbol{w}} = 0 \rightarrow \boldsymbol{w} = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i$$

$$\frac{\partial L}{\partial w_0} = 0 \rightarrow 0 = \sum_{i=1}^{N} \alpha_i y_i$$

3. Substitute the primal variables w and w_0 into the Lagrangian and express in terms of dual variables α_i

$$L = \frac{1}{2}\|\boldsymbol{w}\|_2^2 - \boldsymbol{w}^{\boldsymbol{T}} \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i - w_0 \sum_{i=1}^{N} \alpha_i y_i + \sum_{i=1}^{N} \alpha_i$$

$$= -\frac{1}{2}\boldsymbol{w}^T \boldsymbol{w} + \sum_{i=1}^{N} \alpha_i = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{i'=1}^{N} \alpha_i \alpha_{i'} y_i y_{i'} \boldsymbol{x}_i^T \boldsymbol{x}_{i'}$$

Support Vector Machines: Linearly separable case

$$\min_{w} \frac{1}{2} w^T w, \text{ such that (s.t.) } y_i(w^T x_i + w_0) \geq 1 \text{ for } i = 1,2,\cdots,N$$

4. Maximize the Lagrangian with respect to dual variables (dual problem)

$$\max_{\alpha_i} L = \max_{\alpha_i} \left\{ \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{i'=1}^{N} \alpha_i \alpha_{i'} y_i y_{i'} \boldsymbol{x}_i^T \boldsymbol{x}_{i'} \right\}$$

$$\text{s.t.} \sum_{i=1}^{N} \alpha_i y_i = 0$$

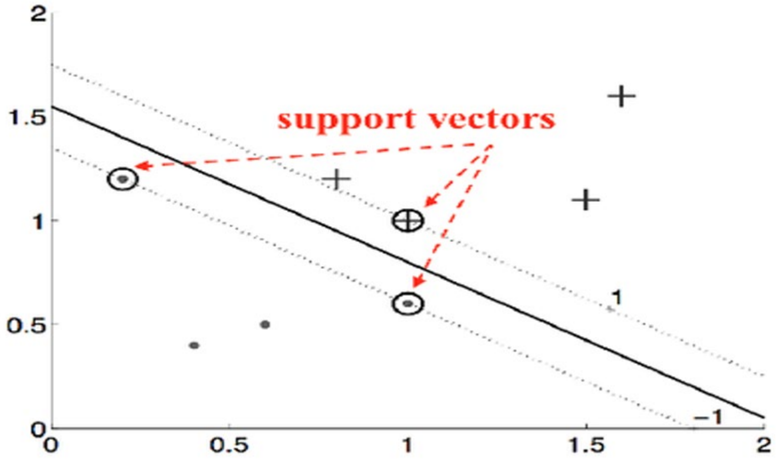
- Solved numerically using quadratic optimization methods
- The dual depends on data size N and no on the data dimensionality D
- Most of the α_i will vanish with $\alpha_i = 0$ only a small percentage
- The set of x_i whose $\alpha_i \neq 0$ are the support vectors

Support Vector Machines: Linearly separable case

$$\min_{w} \frac{1}{2} w^T w, \text{ such that (s.t.) } y_i(w^T x_i + w_0) \geq 1 \text{ for } i = 1,2,\cdots,N$$

5. Recover the solution (for the primal variables) from the dual variables
   - Find w: Substitute α_i from (4) to $w = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i$
   - Find w_0:
     - From $\boldsymbol{w}^T \boldsymbol{x}_i + w_0 = y_i$, where $\boldsymbol{x}_i$ is a support vector, calculate $w_0 = y_i - \boldsymbol{w}^T \boldsymbol{x}_i$
     - For numerical stability, average w_0 values estimated from all support vectors

Support Vector Machines: Linearly separable cases



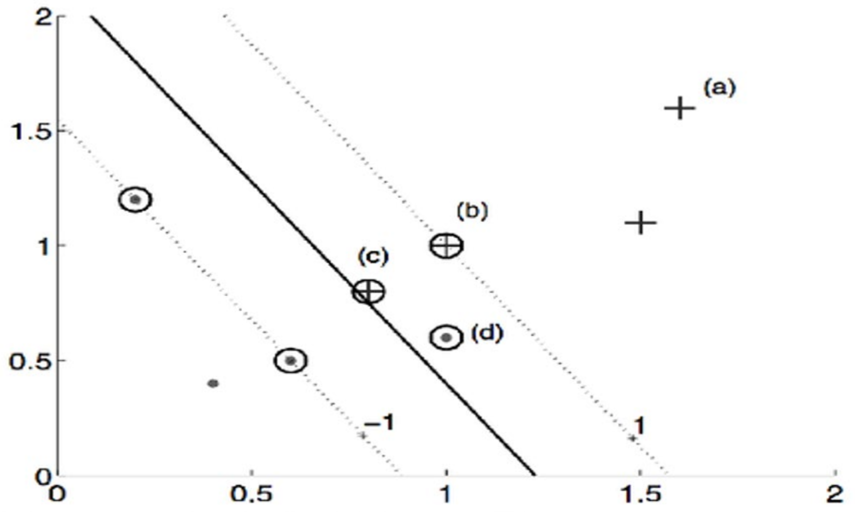
- Sample $\boldsymbol{x}_i$ for which $\alpha_i = 0$
  - Majority of samples
  - Lie away from the hyperplane: $\mathrm{y_i}\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{x}_i + w_0\right) > 1$
  - Have no effect on the hyperplane
- Sample $\boldsymbol{x}_i$ for which $\alpha_i \neq 0$
  - Support Vectors
  - Lie close to the hyperplane: $\mathrm{y_i}\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{x}_i + w_0\right) = 1$
  - Determine the hyperplane

Support Vector Machines: Non-separable case

- If two classes are not linearly separable, we look for the hyperplane that yields the least error
- We define slack variables $\epsilon_i \geq 0$ which represent the deviation from the margin

$$\mathrm{y_i}\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{x}_i + w_0\right) \geq 1 - \epsilon_i$$

- Case (a): Far away from the margin, $\epsilon_i = 0$
- Case (b): On the right side and far from margin, $\epsilon_i = 0$
- Case (c): On the right side, but in the margin, $\epsilon_i > 0$
- Case (d): On the wrong size, $\epsilon_i \geq 1$



Support Vector Machines: non-Linearly separable case

$$\min_{w} \frac{1}{2} w^T w + C \sum_{i=1}^{N} \epsilon_i \text{, such that (s.t.) } y_i(w^T x_i + w_0) \geq 1 - \epsilon_i \text{ and } \epsilon_i > 0 \text{ for } i = 1,2,\cdots,N$$

1. Formulate Lagrangian function (primal problem)

$$L = \frac{1}{2}\|\boldsymbol{w}\|_2^2 - C\sum_{i=1}^{N} \epsilon_i - \sum_{i=1}^{N} \alpha_i \left(y_i(\boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{x}_i + w_0) - 1 + \epsilon_i\right) - \sum_{i=1}^{N} \mu_i \epsilon_i$$

2. Minimize Lagrangian to solve for primal variables w and w_0

$$\frac{\partial L}{\partial \boldsymbol{w}} = 0 \rightarrow \boldsymbol{w} = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i$$

$$\frac{\partial L}{\partial w_0} = 0 \rightarrow 0 = \sum_{i=1}^{N} \alpha_i y_i$$

$$\frac{\partial L}{\partial \epsilon_i} = 0 \rightarrow 0 = C - \alpha_i - \mu_i$$

Support Vector Machines: non-Linearly separable case

$$\min_{w} \frac{1}{2} w^T w + C \sum_{i=1}^{N} \epsilon_i \text{ , such that (s.t.) } y_i(w^T x_i + w_0) \geq 1 - \epsilon_i \text{ and } \epsilon_i > 0 \text{ for } i = 1,2,\cdots,N$$

3. Substitute the primal variables w and w_0 into the Lagrangian and express in terms of dual variables α_i

$$L = \frac{1}{2}\|\boldsymbol{w}\|_2^2 - C\sum_{i=1}^{N}\epsilon_i - \boldsymbol{w}^{\boldsymbol{T}}\sum_{i=1}^{N}\alpha_i y_i \boldsymbol{x}_i - w_0\sum_{i=1}^{N}\alpha_i y_i + \sum_{i=1}^{N}\alpha_i - \sum_{i=1}^{N}\alpha_i\epsilon_i - \sum_{i=1}^{N}\mu_i\epsilon_i$$

$$= \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{i'=1}^{N}\alpha_i\alpha_{i'}y_i y_{i'}\boldsymbol{x}_i^T\boldsymbol{x}_{i'}$$

Scribe Notes: Solve from the beginning of (3) to the end

Support Vector Machines: Linearly separable case

$$\min_{w} \frac{1}{2} w^T w, \text{ such that (s.t.) } y_i(w^T x_i + w_0) \geq 1 \text{ for } i = 1,2,\cdots,N$$

4. Maximize the Lagrangian with respect to dual variables (dual problem)

$$\max_{\alpha_i} L = \max_{\alpha_i} \left\{ \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{i'=1}^{N} \alpha_i \alpha_{i'} y_i y_{i'} \boldsymbol{x}_i^T \boldsymbol{x}_{i'} \right\}$$

$$\text{s.t.} \sum_{i=1}^{N} \alpha_i y_i = 0 \text{ and } 0 \leq \alpha_i \leq C, for\ i = 1,\cdots,N$$

- Solved numerically suing quadratic optimization methods
- The dual depends on data size N and no on the data dimensionality D
- Most of the α_i will vanish with $\alpha_i = 0$ only a small percentage
- The set of x_i whose $\alpha_i > 0$ are the support vectors
  - $0 < \alpha_i < C$: instances lying on the margin
  - $\alpha_i = C$: instances in the margin or misclassified

Takeaways So Far

- SVM aims at finding the hyperplane from which instances have a margin of distance
- Prime and dual problem formulation (Lagrange multiplies)
- Support vectors: instances closest to separating hyperplane
- Linearly separable case: maximize margin of separation between two classes
- Non-separable case: look for the hyperplane that yield the least error (soft erro)
  - Prime: minimizes Lagrangian wrt the primal variables of the problem
  - Dual: maximizes Lagrangian wrt multipliers