

CS195-5 : Introduction to Machine Learning

Lecture 15

Greg Shakhnarovich

October 11, 2006

Announcements

- PS Nectarine due today; PS

Announcements

- PS Nectarine due today; PS ?????? out by Monday.
 - Shorter than usual!
- 10/13: Guest lecture: Meinolf Sellman
 - Optimization, Lagrange multipliers, . . .
- 10/16: no class.
- 10/18: Guest lecture: Chad Jenkins
 - Robot learning, intro to unsupervised and reinforcement learning.

Review

- Ridge regression:

$$\hat{\mathbf{w}}_{ridge} = \underset{\mathbf{w}}{\operatorname{argmax}} - \sum_{i=1}^N (y_i - \mathbf{w}^T \mathbf{x}_i)^2 - \lambda \sum_{j=1}^d w_j^2 = (\lambda \mathbf{I} + \mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}.$$

- Lasso:

$$\hat{\mathbf{w}}_{lasso} = \underset{\mathbf{w}}{\operatorname{argmax}} - \sum_{i=1}^N (y_i - \mathbf{w}^T \mathbf{x}_i)^2 - \lambda \sum_{j=1}^d |w_j|,$$

⇒ no closed-form solution; must solve optimization problem

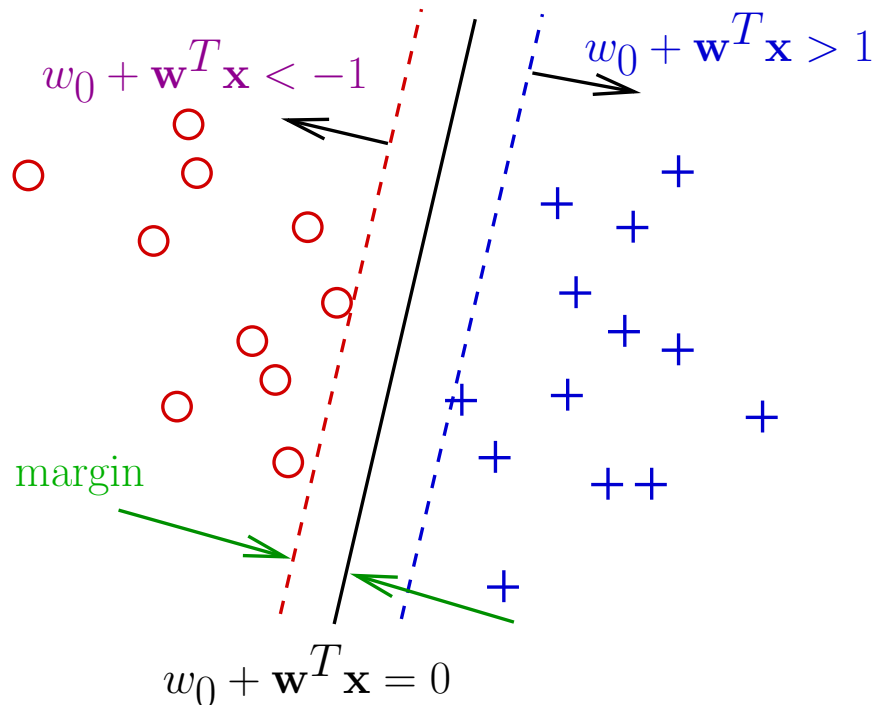
$$\underset{\mathbf{w}: \sum_{j=1}^d |w_j| \leq \beta}{\operatorname{argmax}} - \sum_{i=1}^N (y_i - \mathbf{w}^T \mathbf{x}_i)^2.$$

- Choice of λ or β : by cross-validation.

Plan for today

- Max-margin classification
- Support Vector Machines
 - focus on linearly separable case

The classification margin



- Since the data are separable, we can find \mathbf{w} such that

$$\forall i = 1, \dots, N \quad y_i(w_0 + \mathbf{w}^T \mathbf{x}_i) > 0.$$

- We can even guarantee (by increasing $\|\mathbf{w}\|$ if necessary)

$$\forall i = 1, \dots, N \quad y_i(w_0 + \mathbf{w}^T \mathbf{x}_i) \geq 1.$$

- $\min_i y_i(w_0 + \mathbf{w}^T \mathbf{x}_i)$ is the smallest distance from \mathbf{x}_i to the boundary (half the separation between classes).
- We will refer to it as the *margin*.

Max-margin boundary

- If we just state that we want

$$\hat{\mathbf{w}} = \operatorname{argmax}_{\mathbf{w}} \min_i y_i (w_0 + \mathbf{w}^T \mathbf{x}_i)?$$

we run into the same problem we have seen with LR: when data are separable the margin is unbounded as $\|\mathbf{w}\| \rightarrow \infty$.

- Suppose $y = 1$, and $\|\mathbf{w}\| = 1$. Let $w_0 + \mathbf{w}^T \mathbf{x} = c$. Then,

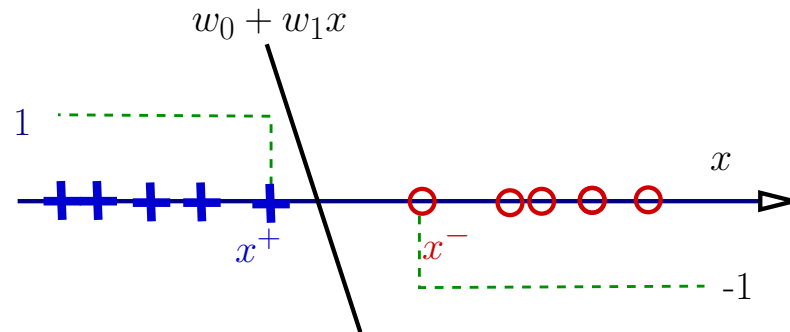
$$\alpha w_0 + (\alpha \cdot \mathbf{w})^T \mathbf{w} = \alpha (w_0 + \mathbf{w}^T \mathbf{x}) = \alpha c,$$

i.e. we can achieve arbitrarily wide margin with the same classification boundary.

- One solution: require $\|\mathbf{w}\| = 1$.

Fixed margin solution

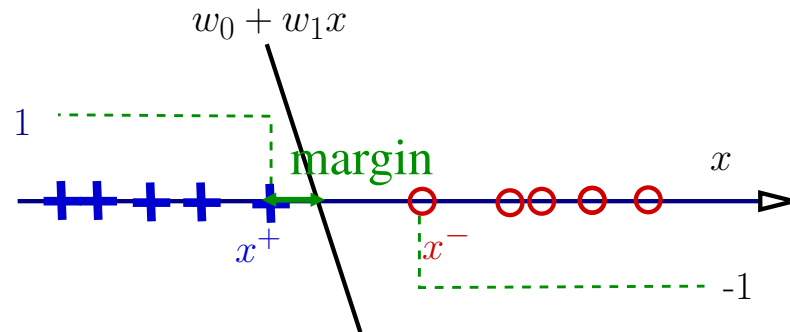
- A more convenient solution: require *fixed* margin of, say, 1.
- Of all \mathbf{w} that achieve such margin, choose the smallest one.
 - This imposes a unique (equivalent) solution!
- The margin constraints, graphically (in 1D):



$$\begin{aligned} 1 \cdot (w_0 + w_1 x_i) - 1 &\geq 0, & y_i &= 1 \\ -1 \cdot (w_0 + w_1 x_i) - 1 &\geq 0, & y_i &= -1. \end{aligned}$$

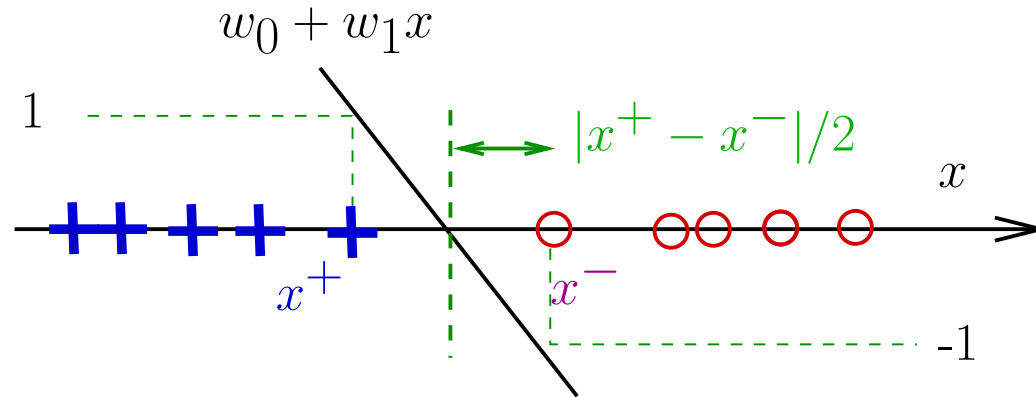
Fixed margin solution

- A more convenient solution: require *fixed* margin of, say, 1.
- Of all \mathbf{w} that achieve such margin, choose the smallest one.
 - This imposes a unique (equivalent) solution!
- The margin constraints, graphically (in 1D):



$$\begin{aligned} 1 \cdot (w_0 + w_1 x_i) - 1 &\geq 0, & y_i &= 1 \\ -1 \cdot (w_0 + w_1 x_i) - 1 &\geq 0, & y_i &= -1. \end{aligned}$$

Margin vs. slope



- Separation is maximal when the line passes through $(x^+ + x^-)/2$.
 - The maximum margin is 1;
- the margin is *inversely proportional* to the slope $|w_1|$;
- The optimal boundary is achieved with margin $1/|w_1|$,

$$|w_1| = 2/|x^+ - x^-|.$$

Margin and regularization

- In general d -dimensional case, we solve the regularization problem:

$$\text{minimize} \quad \|\mathbf{w}\|^2 = \sum_{j=1}^d w_j^2,$$

subject to the margin constraint

$$\forall i, \quad y_i(w_0 + \mathbf{w}^T \mathbf{x}_i) - 1 \geq 0.$$

- This produces margin of exactly 1 (why?)
- Again, the solution is expressed in terms of only a subset of examples.
 - These are the *support vectors*.

Lagrange multipliers

$$\min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|^2 = \frac{1}{2} \sum_{j=1}^d w_j^2,$$

subject to $y_i(w_0 + \mathbf{w}^T \mathbf{x}_i) - 1 \geq 0, \quad i = 1, \dots, N.$

- We will associate with each constraint the loss

$$\max_{\alpha \geq 0} \alpha [1 - y_i(w_0 + \mathbf{w}^T \mathbf{x}_i)] =$$

Lagrange multipliers

$$\min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|^2 = \frac{1}{2} \sum_{j=1}^d w_j^2,$$

subject to $y_i(w_0 + \mathbf{w}^T \mathbf{x}_i) - 1 \geq 0, \quad i = 1, \dots, N.$

- We will associate with each constraint the loss

$$\max_{\alpha \geq 0} \alpha [1 - y_i(w_0 + \mathbf{w}^T \mathbf{x}_i)] = \begin{cases} 0, & \text{if } w_0 + \mathbf{w}^T \mathbf{x}_i - 1 \geq 0, \\ \infty & \text{otherwise (i.e. if constraint violated).} \end{cases}$$

Lagrange multipliers

$$\min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|^2 = \frac{1}{2} \sum_{j=1}^d w_j^2,$$

subject to $y_i(w_0 + \mathbf{w}^T \mathbf{x}_i) - 1 \geq 0, \quad i = 1, \dots, N.$

- We will associate with each constraint the loss

$$\max_{\alpha \geq 0} \alpha [1 - y_i(w_0 + \mathbf{w}^T \mathbf{x}_i)] = \begin{cases} 0, & \text{if } w_0 + \mathbf{w}^T \mathbf{x}_i - 1 \geq 0, \\ \infty & \text{otherwise (i.e. if constraint violated).} \end{cases}$$

- We can reformulate our problem now:

$$\min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^N \max_{\alpha_i \geq 0} \alpha_i [1 - y_i(w_0 + \mathbf{w}^T \mathbf{x}_i)]$$

Max-margin optimization

- We want all the constraint terms to be zero:

$$\begin{aligned} & \min_{\mathbf{w}} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^N \max_{\alpha_i \geq 0} \alpha_i [1 - y_i(w_0 + \mathbf{w}^T \mathbf{x}_i)] \right\} \\ &= \min_{\mathbf{w}} \max_{\{\alpha_i \geq 0\}} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^N \alpha_i [1 - y_i(w_0 + \mathbf{w}^T \mathbf{x}_i)] \right\} \end{aligned}$$

Max-margin optimization

- We want all the constraint terms to be zero:

$$\begin{aligned} & \min_{\mathbf{w}} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^N \max_{\alpha_i \geq 0} \alpha_i [1 - y_i(w_0 + \mathbf{w}^T \mathbf{x}_i)] \right\} \\ &= \min_{\mathbf{w}} \max_{\{\alpha_i \geq 0\}} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^N \alpha_i [1 - y_i(w_0 + \mathbf{w}^T \mathbf{x}_i)] \right\} \\ &= \max_{\{\alpha_i \geq 0\}} \min_{\mathbf{w}} \underbrace{\left\{ \frac{1}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^N \alpha_i [1 - y_i(w_0 + \mathbf{w}^T \mathbf{x}_i)] \right\}}_{J(\mathbf{w}, w_0; \alpha)}. \end{aligned}$$

Max-margin optimization

- We want all the constraint terms to be zero:

$$\begin{aligned} & \min_{\mathbf{w}} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^N \max_{\alpha_i \geq 0} \alpha_i [1 - y_i(w_0 + \mathbf{w}^T \mathbf{x}_i)] \right\} \\ &= \min_{\mathbf{w}} \max_{\{\alpha_i \geq 0\}} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^N \alpha_i [1 - y_i(w_0 + \mathbf{w}^T \mathbf{x}_i)] \right\} \\ &= \max_{\{\alpha_i \geq 0\}} \min_{\mathbf{w}} \underbrace{\left\{ \frac{1}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^N \alpha_i [1 - y_i(w_0 + \mathbf{w}^T \mathbf{x}_i)] \right\}}_{J(\mathbf{w}, w_0; \alpha)}. \end{aligned}$$

- We need to minimize $J(\mathbf{w}, w_0; \alpha)$ for any settings of $\alpha = [\alpha_1, \dots, \alpha_N]^T$.

Strategy for optimization

- We need to find

$$\max_{\{\alpha_i \geq 0\}} \min_{\mathbf{w}} J(\mathbf{w}, w_0; \alpha)$$

- We will first fix α and treat $J(\mathbf{w}, w_0; \alpha)$ as a function of \mathbf{w}, w_0 .
 - Find *functions* $\mathbf{w}(\alpha), w_0(\alpha)$ that attain the minimum.
- Next, treat $J(\mathbf{w}(\alpha), w_0(\alpha); \alpha)$ as a function of α .
 - Find α^* that attain the maximum.
- In the end, the solution is given by $\alpha^*, \mathbf{w}(\alpha^*)$ and $w_0(\alpha^*)$.

Minimizing $J(\mathbf{w}, w_0; \alpha)$ with respect to \mathbf{w}, w_0

- For fixed α we can minimize

$$J(\mathbf{w}, w_0; \alpha) = \frac{1}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^N \alpha_i [1 - y_i(w_0 + \mathbf{w}^T \mathbf{x}_i)]$$

by setting derivatives w.r.t. w_0, \mathbf{w} to zero:

$$\frac{\partial}{\partial \mathbf{w}} J(\mathbf{w}, w_0; \alpha) = \mathbf{w} - \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i = 0,$$

$$\frac{\partial}{\partial w_0} J(\mathbf{w}, w_0; \alpha) = - \sum_{i=1}^N \alpha_i y_i = 0.$$

- Note that the bias term w_0 dropped out but has produced a “global” constraint on α .

Solving for α

$$\mathbf{w}(\alpha) = \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i, \quad \sum_{i=1}^N \alpha_i y_i = 0.$$

- Now can substitute this solution into

$$\begin{aligned} & \max_{\{\alpha_i \geq 0, \sum_i \alpha_i y_i = 0\}} \left\{ \frac{1}{2} \|\mathbf{w}(\alpha)\|^2 + \sum_{i=1}^N \alpha_i [1 - y_i (w_0(\alpha) + \mathbf{w}(\alpha)^T \mathbf{x}_i)] \right\} \\ &= \max_{\{\alpha_i \geq 0, \sum_i \alpha_i y_i = 0\}} \left\{ \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i,j=1}^N \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \right\}. \end{aligned}$$

Max-margin and quadratic programming

- We started by writing down the max-margin problem and arrived at the *dual problem* in α :

$$\max \left\{ \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i,j=1}^N \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \right\}$$

subject to $\sum_{i=1}^N \alpha_i y_i = 0$, $\alpha_i \geq 0$ for all $i = 1, \dots, N$.

- Solving this *quadratic program* yields α^* .
- We substitute α^* back to get \mathbf{w} :

$$\hat{\mathbf{w}} = \mathbf{w}(\alpha^*) = \sum_{i=1}^N \alpha_i^* y_i \mathbf{x}_i$$

Maximum margin decision boundary

$$\hat{\mathbf{w}} = \mathbf{w}(\alpha^*) = \sum_{i=1}^N \alpha_i^* y_i \mathbf{x}_i$$

- Suppose that, under the optimal solution, the margin of \mathbf{x}_i is

$$1 - y_i \hat{\mathbf{w}}^T \mathbf{x}_i > 1.$$

- Then, necessarily, $\alpha_i^* = 0 \Rightarrow$ not a support vector.
- We can then express the direction of the max-margin decision boundary

$$\hat{\mathbf{w}} = \sum_{\alpha_i^* > 0} \alpha_i^* y_i \mathbf{x}_i.$$

- We can compute w_0 by making sure the margin is balanced between the two classes (PS3).

Support vectors

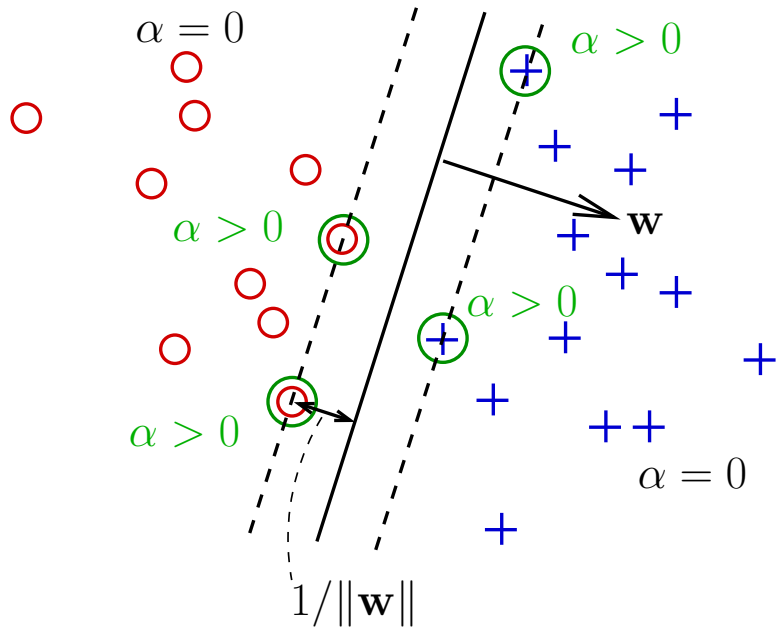
$$\hat{\mathbf{w}} = \sum_{\alpha_i > 0} \alpha_i y_i \mathbf{x}_i.$$

- Given a test example \mathbf{x} , it is classified by

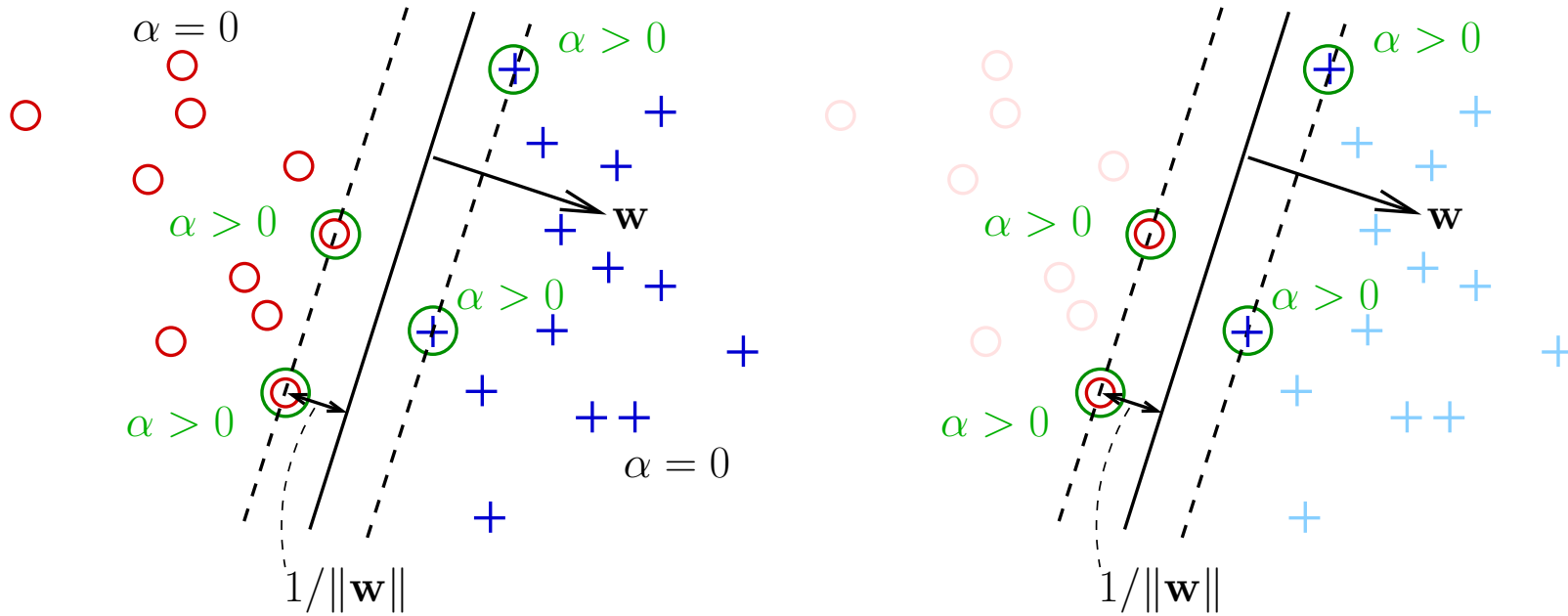
$$\begin{aligned} \hat{y} &= \text{sign}(\hat{w}_0 + \hat{\mathbf{w}}^T \mathbf{x}) \\ &= \text{sign}\left(\hat{w}_0 + \left(\sum_{\alpha_i > 0} \alpha_i y_i \mathbf{x}_i\right)^T \mathbf{x}\right) \\ &= \text{sign}\left(\hat{w}_0 + \sum_{\alpha_i > 0} \alpha_i y_i \mathbf{x}_i^T \mathbf{x}\right) \end{aligned}$$

- The classifier is based on the expansion in terms of dot products of \mathbf{x} with support vectors.

SVM classification



SVM classification



SVM: summary so far

- Assuming linearly separable case, we set up a quadratic program

$$\max \left\{ \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i,j=1}^N \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \right\}$$

$$\text{subject to } \sum_{i=1}^N \alpha_i y_i = 0, \alpha_i \geq 0 \text{ for all } i = 1, \dots, N.$$

- Solving it for α we get the SVM classifier

$$\hat{y} = \text{sign} \left(\hat{w}_0 + \sum_{\alpha_i > 0} \alpha_i y_i \mathbf{x}_i^T \mathbf{x} \right).$$

Non-separable case

- What if the training data are non linearly separable? We can no longer require exact margin constraint.
- We requote the constraints with *slack variables* ξ_i :

$$y_i (w_0 + \mathbf{w}^T \mathbf{x}_i) - 1 + \xi_i \geq 0.$$

- The updated objective:

$$\min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^N \xi_i.$$

- The parameter C determines the penalty paid for violating the exact margin constraints.

Next time

10/13: Optimization.

10/16: no class.

10/18: Robot learning.

10/20: back to SVM: unseparable case, nonlinear classification.