Lecture 12: Support Vector Machines

CS 475: Machine Learning

Raman Arora

March 8, 2017

Review: optimal separating hyperplane

- Decision boundary parametrized as $\mathbf{w} \cdot \mathrm{x} + w_0 = 0$
- "confidence" $= y_i(\mathbf{w} \cdot \mathrm{x}_i + w_0)$
- Distance from the hyperplane $\frac{1}{\|\mathbf{w}\|} y_i(\mathbf{w} \cdot \mathrm{x}_i + w_0)$
- We seek $\mathrm{argmax}_{\mathbf{w},w_0} \left\{ \frac{1}{\|\mathbf{w}\|} \min_i y_i \left(\mathbf{w} \cdot \mathbf{x}_i + w_0\right) \right\}$
- Assuming $\forall i,\ y_i(\mathbf{w} \cdot \mathbf{x}_i + w_0) > 0$, we can rescale $\|\mathbf{w}\|$, w_0 so that

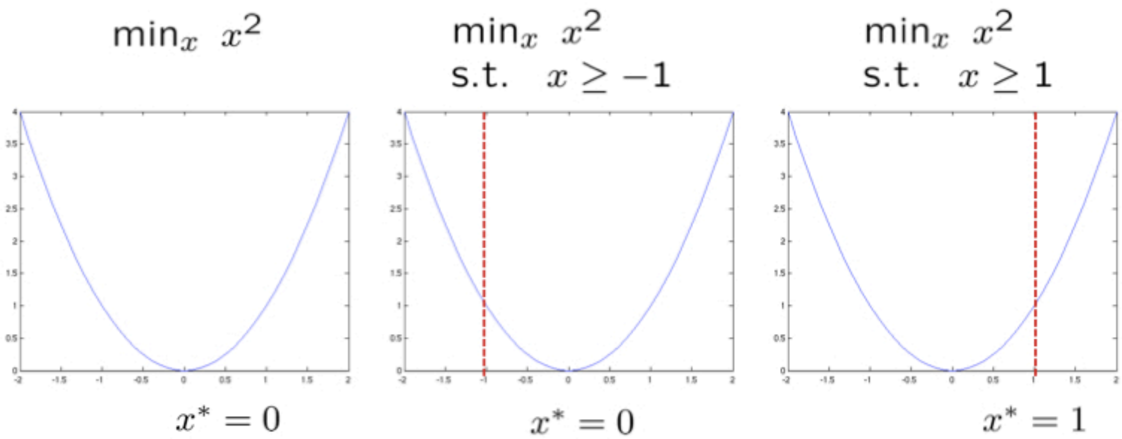
$$\min_i y_i \left(\mathbf{w} \cdot \mathbf{x}_i + w_0\right) = 1.$$

- Then, the optimization becomes:

$$\underset{\mathbf{w},w_0}{\mathrm{argmax}} \quad \frac{1}{\|\mathbf{w}\|} \quad \text{s.t. } y_i \left(\mathbf{w} \cdot \mathbf{x}_i + w_0\right) \geq 1,\ \forall i = 1, \ldots, N.$$

$$\Rightarrow \underset{\mathbf{w}}{\mathrm{argmin}} \quad \|\mathbf{w}\|^2 \quad \text{s.t. } y_i \left(\mathbf{w} \cdot \mathbf{x}_i + w_0\right) \geq 1,\ \forall i = 1, \ldots, N.$$

Review: Constrained optimization



Review: large margin setup

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

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

- We will associate with each constraint the loss

$$\max_{\alpha_i \geq 0} \alpha_i \left[1 - y_i(w_0 + \mathbf{w}\cdot\mathbf{x}_i)\right] = \begin{cases} 0, & \text{if } y_i\left(w_0 + \mathbf{w}\cdot\mathbf{x}_i\right) - 1 \geq 0, \\ \infty & \text{otherwise (constraint violated).} \end{cases}$$

- We can reformulate our problem now:

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

Max-margin optimization

- We want all the constraint terms to be zero:

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

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

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

- Why could we switch $\min$ and $\max$? convexity!

Strategy for optimization

- We need to find

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

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

Minimizing J with respect to $\mathbf{w}, w_0$

- For fixed α we can minimize

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

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

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

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

- Note that the bias term w_0 dropped out but has produced a "global" constraint on $\boldsymbol{\alpha}$.

Solving for α

$$\underbrace{\mathbf{w}(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i\,,}_{\text{later: Representer theorem!}} \qquad \sum_{i=1}^{N} \alpha_i y_i = 0.$$

- Now can (with a bit of algebra) substitute this solution into

$$\max_{\boldsymbol{\alpha}\geq 0, \sum_i \alpha_i y_i = 0} \left\{ \frac{1}{2}\|\mathbf{w}(\boldsymbol{\alpha})\|^2 + \sum_{i=1}^{N} \alpha_i \left[1 - y_i(w_0(\boldsymbol{\alpha}) + \mathbf{w}(\boldsymbol{\alpha})\cdot\mathbf{x}_i)\right] \right\}$$

$$= \max_{\boldsymbol{\alpha}\geq 0, \sum_i \alpha_i y_i = 0} \left\{ \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \right\}.$$

Max-margin and quadratic programming

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

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

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

- Solving this *quadratic program* with linear constraints yields $\boldsymbol{\alpha}^*$.
- We substitute $\boldsymbol{\alpha}^*$ back to get $\mathbf{w}$:

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

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 (distance to the boundary) of a particular $\mathbf{x}_i$ is

$$y_i \left(w_0 + \hat{\mathbf{w}} \cdot \mathbf{x}_i \right) > 1.$$

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

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

- w_0 is set by making the margin equidistant to two classes.

Support vectors

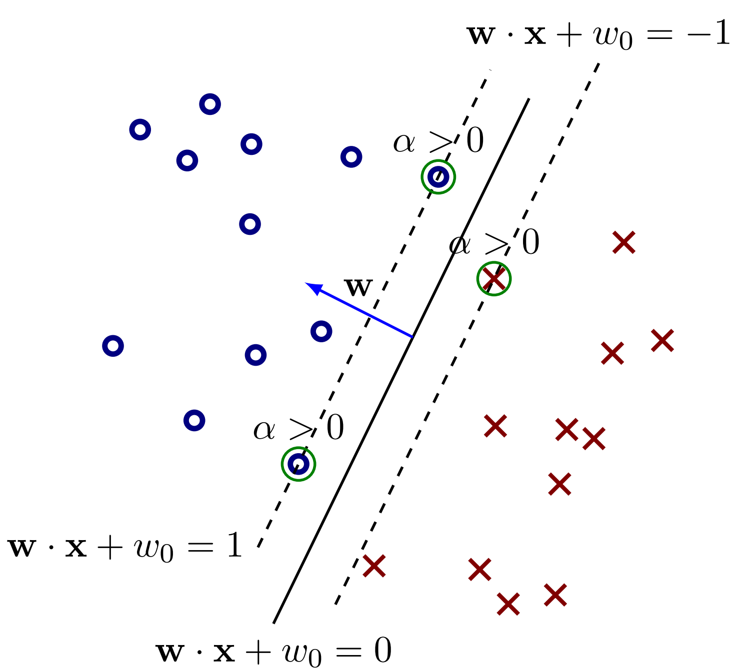
$$\hat{\mathbf{w}} = \sum_{\alpha_i > 0} \alpha_i y_i \mathbf{x}_i.$$

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

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

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

SVM geometry



- Support vectors:

$$\alpha_i > 0$$

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + w_0) = 1$$

- Other examples:

$$\alpha_i = 0$$

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + w_0) > 1$$

Non-separable case

- Not linearly separable data: we can no longer satisfy $y_i\left(\mathbf{w}\cdot\mathbf{x}_i + w_0\right) \geq 1$ for all i.
- Recall the constraint-based terms in separable case:

$$\max_{\boldsymbol{\alpha}\geq 0}\sum_i \alpha_i\left[1 - y_i(w_0 + \mathbf{w}\cdot\mathbf{x}_i)\right]$$

- We can no longer have $\boldsymbol{\alpha}\geq 0$ if constraint violation is unavoidable; would yield $J = \infty$
- We will set maximum penalty on constraint violation:

$$\max_{0\leq\boldsymbol{\alpha}\leq C}\sum_i \alpha_i\left[1 - y_i(w_0 + \mathbf{w}\cdot\mathbf{x}_i)\right]$$

Slack variables

- We introduce *slack variables* to satisfy margin constraints

$$y_i\left(w_0 + \mathbf{w}\cdot\mathbf{x}_i\right) - 1 + \xi_i \;\geq\; 0, \qquad \xi_i \;\geq\; 0.$$

- We want ξ_i to capture the *minimum* amount we need to fix:

$$\xi_i \;=\; \max\left\{0,\, 1 - y_i\left(w_0 + \mathbf{w}\cdot\mathbf{x}_i\right)\right\}$$

  note: ξ_i is really a function of $\mathbf{w}$
- Our objective now:

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

Non-separable case: solution

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

- We can solve this using Lagrange multipliers
  - Introduce additional multipliers for the $\xi \geq 0$.
- The resulting dual problem:

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

subject to $\sum_{i=1}^{N} \alpha_i y_i = 0, \; 0 \leq \boldsymbol{\alpha} \leq C.$

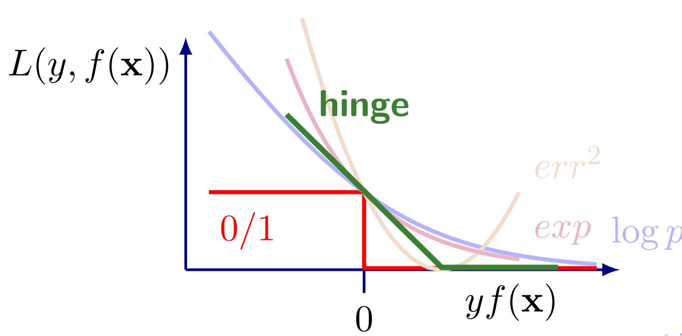
Loss in SVM

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

- L_2-regularized loss, measured as

$$\sum_{i=1}^{N} \xi_i \; = \sum_{i=1}^{N} \max\left\{0, 1 - y_i(w_0 + \mathbf{w}\cdot\mathbf{x}_i)\right\}$$

- This surrogate loss is known as *hinge loss*



Solving SVM in the primal

- Setting $\lambda = 2/C$ we get

$$\text{primal:} \qquad \min_{\mathbf{w}} \frac{\lambda}{2}\|\mathbf{w}\|^2 \; + \; \sum_{i=1}^{N} \max\left\{0, 1 - y_i \mathbf{w} \cdot \mathbf{x}_i\right\}$$

- Traditional tactic: write the dual, solve using QP
- Alternative: optimize the primal directly using gradient descent
- Problem: hinge loss is not differentiable at $y\mathbf{w} \cdot \mathbf{x} = 1$
- Solution: *sub*gradient descent