Lecture #19: 11-20-18

Recall: When confronted with the constrained optimization problem

$$\min_{x} f(x) \text{ subject to } \overset{h}{\cancel{g}}(x) \geq 0, \quad \overset{h}{\cancel{g}} = (\overset{h}{\cancel{g}}_1, \dots, \overset{h}{\cancel{g}}_n), \qquad \underline{\text{PRIMAL}}$$

Solving the dual problem provides a lower bound:

$$\max_{\lambda \geq 0} g(\lambda), \quad \text{where} \quad g(\lambda) = \min_{x \in \mathbb{R}} f_\lambda(x) \qquad \underline{\text{DUAL}}$$

$$= \min_{x \in \mathbb{R}} f(x) - \sum_{i=1}^{n} \lambda_i \overset{h}{\cancel{g}}_i(x).$$

- By lower bound, we mean $\max_{\lambda \geq 0} g(\lambda) \leq \min_{x \in \mathbb{R}^d} f(x)$.

The duality gap $= \left(\min_{x \in \mathbb{R}^d} f(x) \right) - \max_{\lambda \geq 0} g(\lambda)$; when this gap $= 0$, solving the dual problem is equivalent to solving the primal problem.

//

Returning to SVM, recall that the hard margin SVM formulation seeks to learn a separating hyperplane $\{x \mid x^T w + b\}$ s.t. all points with one label are on one side, all points with the other label on the opposite side. This amounts to the following optimization problem:

- $\{x_i\}_{i=1}^n \subset \mathbb{R}^d$ data points
- $\{y_i\}_{i=1}^n$, $y_i \in \{-1,1\}$ labels
- $\underset{\mathbb{R}^d}{\underset{\uparrow}{w}}, \underset{\mathbb{R}}{\underset{\uparrow}{b}}$ hyperplane parameters

$$(w^*, b^*) = \arg\min_{w} \|w\|_2^2 \quad \text{subject to} \quad y_i(w^T x_i + b) \geq 1.$$

This is the primal formulation. Setting
$$f(w,b) = \|w\|_2^2$$
$$h_i(w,b) = y_i(w^T x_i + b) - 1,$$

this optimization problem is of the form $\arg\min_{(w,b)} f(x,b)$ s.t. $h_i(w,b) \geq 0$ for each $i=1..n$.

We can thus pass to the dual problem: $\arg\max_{\lambda \geq 0} g(\lambda)$, where

$$g(\lambda) = \min_{(w,b)} f(w,b) - \sum_{i=1}^n \lambda_i h_i(w,b)$$
$$= \min_{(w,b)} \|w\|_2^2 - \sum_{i=1}^n \lambda_i [y_i(w^T x_i + b) - 1].$$

For fixed λ, this is easily minimized in (w,b):

$$\|w\|_2^2 - \sum_{i=1}^n \lambda_i \left(y_i(w^T x_i + b) - 1 \right)$$

$$= \|w\|_2^2 - \sum_{i=1}^n \lambda_i y_i (w^T x_i + b) + \sum_{i=1}^n \lambda_i .$$

Note: $-\sum_{i=1}^n \lambda_i y_i (w^T x_i + b) = -\sum_{i=1}^n \lambda_i y_i w^T x_i - b \sum_{i=1}^n \lambda_i y_i .$

So, if we send $b \to \infty$, the function we are trying to minimize $\to -\infty$, unless $\sum_{i=1}^n \lambda_i y_i = 0$. This is the case we thus focus on.

In the case this occurs, we seek to minimize

$$\|w\|_2^2 - \sum_{i=1}^n \lambda_i y_i w^T x_i + \sum_{i=1}^n \lambda_i$$

Exercise: This minimum occurs when $w = \frac{1}{2} \sum_{i=1}^n \lambda_i y_i x_i$.

Thus, $g(\lambda) = \begin{cases} -\frac{1}{4} \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j y_i y_j x_i^T x_j + \sum_{i=1}^n \lambda_i, & \text{if } \sum_{i=1}^n y_i \lambda_i = 0 \\ -\infty, & \text{if } \sum_{i=1}^n y_i \lambda_i \neq 0 \end{cases}$

- So, solving the dual problem amounts to maximizing this $g(\lambda)$ over $\lambda \geq 0$:

$$\max_{\lambda \geq 0} g(\lambda) = \max_{\lambda \geq 0} -\frac{1}{4} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j x_i^T x_j + \sum_{i=1}^{n} \lambda_i$$

$$\text{subject to } \sum_{i=1}^{n} \lambda_i y_i = 0.$$

This is a quadratic problem in λ, and is easily solved with convex programming.

- Given the optimal $\lambda^* = (\lambda_1^*, \ldots, \lambda_n^*)$, w can be found by minimizing the unconstrained problem

$$\|w\|^2 - \sum_{i=1}^{n} \lambda_i^* y_i \left((w^T x_i + b) - 1 \right)$$

- To get b, we set

$$b = \max_{i \text{ s.t. } y_i = 1} (1 - w^T x_i) = \min_{i \text{ s.t. } y_i = -1} (-1 - w^T x_i),$$

which has the effect of maximizing the margin.

- What about for the soft margin SVM

$$(w^*, b^*) = \arg\min_{(w,b)} \|w\|_2^2 - \alpha \sum_{i=1}^{n} \max(0, 1 - y_i (w^T x_i + b)).$$

- This is unconstrained... how to cast this in the dual formulation?
- We introduce so-called <u>slack variables</u>:

$$\min_{\substack{(w,b,\xi) \\ w\in\mathbb{R}^d \\ b\in\mathbb{R} \\ \xi\in\mathbb{R}^n}} F(w,b,\xi) = \|w\|_2^2 + C\|\xi\|_2^2, \quad \text{subject to } y_i(w^T x_i + b) \geq 1-\xi_i \text{ and } \xi_i \geq 0.$$

<u>Exercise</u>: Prove this is equivalent to the soft-margin SVM formulation.

- So, there are n constraints of the form $y_i(w^T x_i + b) \geq 1-\xi_i$
  n constraints of the form $\xi_i \geq 0$
- So, we seek in the dual problem to maximize $g(\lambda,\mu)$ where

$$g(\lambda,\mu) = \min_{w,b,\xi}\left(\|w\|_2^2 + C\|\xi\|_2^2 - \sum_{i=1}^n \lambda_i\left[y_i(w^T x_i + b) - 1 + \xi_i\right] - \sum_{i=1}^n \mu_i \xi_i\right)$$

subject to $\lambda_i \geq 0$, $\mu_i \geq 0$.

- As in the case of the hard margin SVM, the minimum $= -\infty$ unless $\sum_{i=1}^n y_i\lambda_i = 0$, and it is an <u>exercise</u> to show

$$g(\lambda,\mu) = \begin{cases} -\frac{1}{4}\sum_{i=1}^n\sum_{j=1}^n \lambda_i\lambda_j y_i y_j x_i^T x_j + \sum_{i=1}^n \lambda_i - \frac{1}{4C}\|\lambda+\mu\|_2^2, & \text{if } \sum_{i=1}^n \lambda_i y_i = 0 \\ -\infty, & \text{else} \end{cases}$$

Solving for w, b, we get $$w = \frac{1}{2}\sum_{i=1}^{n} \lambda_i y_i x_i$$

$$b = y_i(1-\xi_i) - w^T x_i, \text{ any } i \text{ with } \xi_i > 0$$

<u>Next time</u>: Kernel SVM for non-linear boundaries.