F. SML: Lecture 14

- We are now in a position to prove a learning estimate for SVM.

Defn (leave-one-out error): Let h_S be the hypothesis returned by an algorithm $\mathcal{A}$ when trained on a finite sample S. The leave-one-out error of $\mathcal{A}$ on S is

$$\hat{R}_{LOO}(\mathcal{A}) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{h_{S\setminus\{x_i\}}(x_i)\neq y_i}$$

- In other words, for each x_i, we train on $S\setminus\{x_i\}$, then compute the error of predicting on x_i. We then average these errors.

Lemma: The average LOO error for samples of size $m \geq 2$ is an unbiased estimate of the average generalization error for samples of size $m-1$:

$$\mathop{\mathbb{E}}_{S\sim D^m}\left(\hat{R}_{LOO}(\mathcal{A})\right) = \mathop{\mathbb{E}}_{S'\sim D^{m-1}}\left(R(h_{S'})\right).$$

Proof: By linearity of expectation, we compute

$$\mathop{\mathbb{E}}_{S\sim D^m}\left(\hat{R}_{LOO}(\mathcal{A})\right) = \frac{1}{m}\sum_{i=1}^{m} \mathop{\mathbb{E}}_{S\sim D^m}\left(\mathbb{1}_{h_{S\setminus\{x_i\}}(x_i)\neq y_i}\right)$$

$$\xrightarrow{\text{iid samples}} = \mathop{\mathbb{E}}_{S\sim D^m}\left(\mathbb{1}_{h_{S\setminus\{x_1\}}(x_1)\neq y_1}\right)$$

$$= \mathop{\mathbb{E}}_{\substack{S'\sim D^{m-1}\\ x_1\sim D}}\left(\mathbb{1}_{h_{S'}(x_1)\neq y_1}\right)$$

$$= \mathop{\mathbb{E}}_{S' \sim D^{m-1}} \left(\mathop{\mathbb{E}}_{x_1 \sim D} \left(1_{h_{S'}(x_1) \neq y_1} \right) \right)$$

$$= \mathop{\mathbb{E}}_{S' \sim D^{m-1}} \left(R(h_{S'}) \right) . \blacksquare$$

Theorem (learning bound for linear SVM): Let h_S be the hypothesis returned by SVM for a sample S, and let $N_{SV}(S)$ be the number of support vectors that define h_S.

Then: $$\mathop{\mathbb{E}}_{S \sim D^m} \left(R(h_S) \right) \leq \mathop{\mathbb{E}}_{S \sim D^{m+1}} \left(\frac{N_{SV}(S)}{m+1} \right).$$

Proof: Let S be a linearly separable sample of size $m+1$. If x is not a support vector, then removing it from the sample does not change the SVM solution. Hence, $h_{S \setminus \{x\}} = h_S$ and $h_{S \setminus \{x\}}$ correctly classifies x. Hence, if $h_{S \setminus \{x\}}$ misclassifies x, then x must be a support vector. Thus,

$$\hat{R}_{LOO}(SVM) \leq \frac{N_{SV}(S)}{m+1}$$

$$\Rightarrow \mathop{\mathbb{E}}_{S \sim D^m} \left(\hat{R}_{LOO}(SVM) \right) \leq \mathop{\mathbb{E}}_{S \sim D^m} \left(\frac{N_{SV}(S)}{m+1} \right)$$

$$\| \qquad\qquad\qquad\qquad \|$$

$$\mathop{\mathbb{E}}_{S \sim D^m} \left(R(h_S) \right) . \blacksquare$$

- So far, we have assumed that the data is linearly separable, i.e. that there exists some $(w,b) \in \mathbb{R}^{D+1}$ such that $y_i(w \cdot x_i + b) \geq 1 \quad \forall i = 1, \dots, m$.

- This is not really very realistic, so we consider relaxing these conditions as follows. For each $i = 1, \dots, m$, suppose $\exists\, \xi_i \geq 0$ s.t.

$$y_i[w \cdot x_i + b] \geq 1 - \xi_i$$

We call the $\{\xi_i\}_{i=1}^m$ slack variables.

- If $\xi_i > 0$, this means x_i is on the "wrong side" of the hyperplane, and may be considered an outlier.

- We would like a large margin classifier still, but we would also like the total amount of slack (e.g. $\sum_{i=1}^m \xi_i$) to be small. These are in tension.

- Consider the following constrained optimization problem: for $\lambda > 0$, $\rho \geq 1$,

$$\min_{\substack{(w,b) \in \mathbb{R}^{D+1} \\ \xi \in \mathbb{R}^m}} \frac{1}{2}\|w\|_2^2 + \underbrace{\lambda \sum_{i=1}^m \xi_i^\rho}_{\lambda \|\xi\|_\rho^\rho} \quad \text{subject to} \quad y_i(w \cdot x_i + b) \geq 1 - \xi_i \quad \text{and} \quad \xi_i \geq 0 \quad \forall i = 1, \dots, m \qquad (\star)$$

- The ρ-norms most commonly considered for the slack penalty are $\rho = 1$ and $\rho = 2$. In what follows, we will analyze the $\rho = 1$ case.

Remark: This optimization $(\star)$ is convex, since the objective function is convex, and the constraints are affine.

- As in the separable case, we may introduce the Lagrangian and examine the KKT conditions.

- We have $2m$ constraints. Let $\{\alpha_i\}_{i=1}^m$ be the multipliers associated to $y_i(w\cdot x_i+b)\geq 1-\xi_i$ and let $\{\beta_i\}_{i=1}^m$ be the multipliers associated to $\xi_i \geq 0$. This yields a Lagrangian

$$\mathcal{L}(w,b,\xi,\alpha,\beta) = \frac{1}{2}\|w\|_2^2 + \lambda\sum_{i=1}^m \xi_i - \sum_{i=1}^m \alpha_i\left(y_i(w\cdot x_i+b)-1+\xi_i\right) - \sum_{i=1}^m \beta_i\xi_i$$

- Setting $\nabla_w\mathcal{L}$, $\nabla_b\mathcal{L}$, $\nabla_\xi\mathcal{L}$ equal to 0 yields

- $\nabla_w\mathcal{L} = w - \sum_{i=1}^m \alpha_i y_i x_i = 0 \Rightarrow w = \sum_{i=1}^m \alpha_i y_i x_i$

- $\nabla_b\mathcal{L} = -\sum_{i=1}^m \alpha_i y_i = 0 \Rightarrow 0 = \sum_{i=1}^m \alpha_i y_i$

- $\nabla_\xi\mathcal{L} = \lambda - (\alpha_i+\beta_i) = 0 \Rightarrow \lambda = \alpha_i+\beta_i \quad \forall i$

- Moreover, $\forall i \ \alpha_i\left(y_i(w\cdot x_i+b)-1+\xi_i\right)=0 \Rightarrow \alpha_i = 0$ or $y_i(w\cdot x_i+b) = 1-\xi_i$

- $\forall i \ \beta_i\xi_i = 0 \Rightarrow \beta_i = 0$ or $\xi_i = 0$.

- As before, x_i contributes to $w \Leftrightarrow \alpha_i \neq 0$. But, $\alpha_i \neq 0 \Leftrightarrow y_i(w\cdot x_i+b) = 1-\xi_i$

- If $\xi_i = 0$, then $y_i(w\cdot x_i+b)=1$ and x_i is on a marginal hyperplane, as in the case when the data is linearly separable. Otherwise if $\xi_i > 0$, then $\beta = 0 \Rightarrow \alpha_i = \lambda$

- To summarize, either support vectors lie on the marginal hyperplanes ($\xi_i = 0$) or else they are outliers and $\alpha_i = \lambda$ ($\xi_i > 0$).