Mehryar Mohri
Advanced Machine Learning 2015
Courant Institute of Mathematical Sciences
Homework assignment 1
February 27, 2015
Due: March 16, 2015

A. Learning kernels

Consider the learning kernel optimization based on SVM:

$$\min_{\mu\in\Delta_\infty} \max_{\boldsymbol{\alpha}} \; 2\,\boldsymbol{\alpha}^\top \mathbf{1} - \boldsymbol{\alpha}^\top \mathbf{Y}^\top \mathbf{K}_{\boldsymbol{\mu}} \mathbf{Y} \boldsymbol{\alpha}$$
$$\text{subject to: } \mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{C} \wedge \boldsymbol{\alpha}^\top \mathbf{y} = 0,$$

where $\Delta_\infty = \{\boldsymbol{\mu} \colon \|\boldsymbol{\mu}\|_\infty \leq 1 \ \wedge \boldsymbol{\mu} \geq 0\}$ and where for the rest the notation is the one used in the lecture slides. Show that its solution coincides with the SVM solution for the uniform combination kernel.

Solution: Let $f(\boldsymbol{\alpha}, \boldsymbol{\mu}) = 2\boldsymbol{\alpha}\mathbf{1} - \sum_{j=1}^n \mu_j \boldsymbol{\alpha}^\top \mathbf{Y}^\top \mathbf{K}_j \mathbf{Y} \boldsymbol{\alpha}$, $C_\alpha = \{\boldsymbol{\alpha} | 0 \leq \boldsymbol{\alpha} \leq \mathbf{C} \ \wedge \ \boldsymbol{\alpha}^\top \mathbf{y} = 0\}$. The function f is a linear in $\boldsymbol{\mu}$ and therefore convex. Since $\mathbf{K}_j$ is positive definite it also follows that f is a concave function of $\boldsymbol{\alpha}$. Furthermore, C_α and Δ_∞ are both compact convex sets. By the minimax theorem we therefore have $\min_{\boldsymbol{\mu}\in\Delta_\infty} \max_{\boldsymbol{\alpha}\in C_\alpha} f(\boldsymbol{\alpha}, \boldsymbol{\mu}) = \max_{\boldsymbol{\alpha}\in C_\alpha} \min_{\boldsymbol{\mu}\in\Delta_\infty} f(\boldsymbol{\alpha}, \boldsymbol{\mu})$. Using the definition of f we see

$$\begin{aligned}\min_{\boldsymbol{\mu}\in\Delta_\infty} f(\boldsymbol{\alpha}, \boldsymbol{\mu}) &= \min_{\boldsymbol{\mu}\in\Delta_\infty} 2\boldsymbol{\alpha}\mathbf{1} - \sum_{j=1}^n \mu_j \boldsymbol{\alpha}^\top \mathbf{Y}^\top \mathbf{K} \mathbf{Y} \boldsymbol{\alpha} \\ &= 2\boldsymbol{\alpha}\mathbf{1} - \max_{\boldsymbol{\mu}\in\Delta_\infty} \sum_{j=1}^n \mu_j \boldsymbol{\alpha}^\top \mathbf{Y}^\top \mathbf{K}_j \mathbf{Y} \boldsymbol{\alpha}.\end{aligned}$$

Since $\boldsymbol{\alpha}^\top \mathbf{Y}^\top \mathbf{K}_j \mathbf{Y} \boldsymbol{\alpha} \geq 0$ for all j, the above maximum is attained at $\mu_j = 1 \forall j$. Therefore, the solution coincides with the uniform combination kernel.

B. Cross-Validation

The objective of this problem is to derive a learning bound for cross-validation comparing its performance to that of SRM. Let $(H_k)_{k\in\mathbb{N}}$ be a countable sequence of hypothesis sets with increasing complexities.

The cross-validation (CV) solution is obtained as follows. Suppose the learner receives an i.i.d. sample S of size $m \geq 1$. He randomly divides S into a sample S_1 of size $(1-\alpha)m$ and a sample S_2 of size αm, where α is in $(0,1)$, with α typically small. S_1 is used for training, S_2 for validation. For any $k \in \mathbb{N}$, let $\widehat{h}_k$ denote the solution of ERM run on S_1 using hypothesis set H_k. The learner then uses sample S_2 to return the CV solution $f_{\text{CV}} = \operatorname{argmin}_{k\in\mathbb{N}} \widehat{R}_{S_2}(\widehat{h}_k)$.

1. Prove the following inequality:

$$\Pr\left[\sup_{k\geq 1}\left|R(\widehat{h}_k) - \widehat{R}_{S_2}(\widehat{h}_k)\right| > \epsilon + \sqrt{\frac{\log k}{\alpha m}}\right] \leq 4e^{-2\alpha m\epsilon^2}.$$

Solution: By the union bound we have

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{k\geq 1}\left|R(\widehat{h}_k) - \widehat{R}_{S_2}(\widehat{h}_k)\right| > \epsilon + \sqrt{\frac{\log k}{\alpha m}}\Big)\\
&\leq \sum_{k=1}^{\infty}\mathbb{P}\Big(\left|R(\widehat{h}_k) - \widehat{R}_{S_2}(\widehat{h}_k)\right| > \epsilon + \sqrt{\frac{\log k}{\alpha m}}\Big)\\
&= \sum_{k=1}^{\infty}\mathrm{E}\left[\mathbb{P}\Big(\left|R(\widehat{h}_k) - \widehat{R}_{S_2}(\widehat{h}_k)\right| > \epsilon + \sqrt{\frac{\log k}{\alpha m}}|S_1\Big)\right]. \qquad (1)
\end{aligned}$$

Hypothesis $\widehat{h}_k$ is fixed conditioned on S_1. Furthermore, sample S_2 is independent from sample S_1 therefore, by Hoeffding's inequality we can bound the conditional probability by

$$\begin{aligned}
\mathbb{P}\Big(\left|R(\widehat{h}_k) - \widehat{R}_{S_2}(\widehat{h}_k)\right| > \epsilon + \sqrt{\frac{\log k}{\alpha m}}|S_1\Big) &\leq 2e^{-2\alpha m\left(\epsilon+\sqrt{\frac{\log k}{\alpha m}}\right)^2}.\\
&\leq e^{-2\alpha m\epsilon^2 - 2\log k}\\
&= \frac{1}{k^2}e^{-2\alpha m\epsilon^2}.
\end{aligned}$$

Replacing this bound in (1) and summing over k we obtain

$$\mathbb{P}\Big(\sup_{k\geq 1}\left|R(\widehat{h}_k) - \widehat{R}_{S_2}(\widehat{h}_k)\right| > \epsilon + \sqrt{\frac{\log k}{\alpha m}}\Big) \leq \frac{\pi^2}{3}e^{2\alpha m\epsilon^2} < 4e^{2\alpha m\epsilon^2}.$$

2. Let $R(f_{\text{SRM}}, S_1)$ be the generalization error of the SRM solution using a sample S_1 of size $(1-\alpha m)$ and $R(f_{\text{CV}}, S)$ the generalization error of the cross-validation solution using a sample S of size m. Use the previous question to prove that for any $\delta > 0$, with probability at least $1-\delta$ the following holds:

$$R(f_{\text{CV}}, S) - R(f_{\text{SRM}}, S_1) \leq 2\sqrt{\frac{\log \frac{4}{\delta}}{2\alpha m}} + 2\sqrt{\frac{\log \max(k(f_{\text{CV}}), k(f_{\text{SRM}}))}{\alpha m}},$$

where, as for the notation used in class, for any h, $k(h)$ denotes the smallest index of a hypothesis set contained h. Comment on the bound derived: point out both the usefulness its suggests for CV and its possible drawback in some bad cases.

Solution: In view of the previous bound we know that with probability at least $1-\delta$ the following holds:

$$\begin{aligned}
R(f_{CV}, S) &\leq \widehat{R}_{S_2}(f_{CV}) + \sqrt{\frac{\log \frac{4}{\delta}}{2\alpha m}} + \sqrt{\frac{\log(k(f_{CV}))}{\alpha m}} \\
&\leq \widehat{R}_{S_2}(f_{SRM}) + \sqrt{\frac{\log \frac{4}{\delta}}{2\alpha m}} + \sqrt{\frac{\log(k(f_{CV}))}{\alpha m}} \\
&\leq R(f_{SRM}, S_1) + 2\sqrt{\frac{\log \frac{4}{\delta}}{2\alpha m}} + \sqrt{\frac{\log(k(f_{CV}))}{\alpha m}} + \sqrt{\frac{\log(k(f_{SRM}))}{\alpha m}} \\
&\leq R(f_{SRM}, S_1) + 2\sqrt{\frac{\log \frac{4}{\delta}}{2\alpha m}} + 2\sqrt{\frac{\log(\max(k(f_{CV}), k(f_{SRM}))}{\alpha m}},
\end{aligned}$$

where we have used the definition of f_{CV} as a minimizer for the second inequality and the third inequality follows again from the previuosly derived bound. The bound tells us that the error of the classifier obtained through cross validation will be close to the value of the classifier obtained through SRM. However, we are interested in the SRM solution of training on m points and the error of training on $(1-\alpha)m$ points could be, in some bad cases, much worse than its error on m points.

3. Suppose that for any k, $\widehat{R}_{S_1}(\widehat{h}_{k+1}) < \widehat{R}_{S_1}(\widehat{h}_k)$ for all k such that $\widehat{R}_{S_1}(\widehat{h}_k) > 0$ and $\widehat{R}_{S_1}(\widehat{h}_{k+1}) \leq \widehat{R}_{S_1}(\widehat{h}_k)$ otherwise. Show that we can then restrict the analysis to H_ks with $k \leq m+1$ and give a more explicit guarantee similar to that of the previous question.

Solution: From the fact that $\widehat{R}_{S_1}(\widehat{h}_{k+1}) < \widehat{R}_{S_1}(\widehat{h}_k)$ it follows that $\widehat{R}_{S_1}(\widehat{h}_{k+1}) \leq \widehat{R}_{S_1}(\widehat{h}_k) - \frac{1}{m}$. Thus, by induction, $\widehat{R}_{S_1}(\widehat{h}_n) = 0$ for all $n \geq m+1$. This implies that $\widehat{h}_n = \widehat{h}_{m+1}$ for all $n \geq m+1$ and therefore we may assume that $k(f_{CV}) \leq m+1$ and since the complexity of H_k increases with k we also have $k(f_{SRM}) \leq m+1$. In view of this, we obtain the more explicit bound

$$R(f_{CV}, S) - R(f_{SRM}, S_1) \leq 2\sqrt{\frac{\log(\frac{4}{\delta})}{2\alpha m}} + 2\sqrt{\frac{\log(m+1)}{\alpha m}}$$

C. CRF

In class, we discussed the learning algorithms for CRF in the case of bigram features.

1. Write the expression of the features in the case of n-grams (arbitrary $n \geq 1$). *Solution:* Following the notation used in class we can write the expression for n grams as

   $$\mathbf{\Phi}(\mathbf{x}, \mathbf{y}) = (\phi(\mathbf{x}, 1, y_1), \ldots, \phi(\mathbf{x}, 1, y_{k-n}, \ldots, y_k), \ldots, \phi(\mathbf{x}, l, y_{l-n}, \ldots, y_l)).$$

2. Describe explicitly the key graph-based algorithms in the case of n-grams. What is the running-time complexity of these algorithms? *Solution:* In order to find

   $$\operatorname*{argmax}_{\mathbf{y} \in \Delta^l} \mathbf{w} \cdot \mathbf{\Phi}(x, y) = \operatorname*{argmax}_{\mathbf{y} \in \Delta^l} \mathbf{w} \cdot \phi(\mathbf{x}, k, y_{k-n}, \ldots, y_k)$$

   we use again a single-source shortest-distance algorithm. In the case of n-grams the corresponding graph has $r^n(l-(n-1)) + r\frac{r^{n-1}-1}{r-1} + r$. In order to see this, notice that nodes in column k are be of the form:

   $$(y_{k-n+1}, \ldots, y_k, k)$$

   And there will be an edge between nodes $(y'_{k-n}, \ldots, y'_{k-1}, k-1)$ and $(y_{k-n+1}, \ldots, y_k)$. If an only if

   $$y'_i = y_i \quad \forall\, i \in \{k-n+1, \ldots, k-1\}.$$

   This edge will correspond to the dot product $\mathbf{w} \cdot \phi(\mathbf{x}, k, y_{k-n}, \ldots, y_k)$. It is easy to verify that for $k \geq (n-1)$ the number of nodes in each

column is r^{n-1}. Moreover each node in column $k-1$ can be matched with exactly r nodes in column k. Therefore each column has r^n edges. A similar argument shows that for $k \leq (n-1)$ each column has r^k edges and the 0-th column has r edges. Therefore the whole graph has

$$(l-(n-1))r^n + \sum_{k=1}^{n-1} r^k + r = (l-(n-1))r^n + r\frac{r^{n-1}-1}{r-1} + r$$

edges. Using a linear time complexity algorithm for finding a shortest path it follows that the complexity of our algorithm is in $O(lr^n)$.