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.

**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 $\alpha m$, where $\alpha$ is in $(0,1)$, with $\alpha$ 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 $f_{\mathrm{CV}} = \operatorname{argmin}_{k \in \mathbb{N}} \widehat{R}_{S_2}(\widehat{h}_k)$.

1. Prove the following inequality:

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

2. Let $R(f_{\mathrm{SRM}}, S_1)$ be the generalization error of the SRM solution using a sample $S_1$ of size $(1-\alpha m)$ and $R(f_{\mathrm{CV}}, S)$ 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.

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 $H_k$s with $k \leq m + 1$ and give a more explicit guarantee similar to that of the previous question.

**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 $n \geq 1$).

2. Describe explicitly the key graph-based algorithms in the case of $n$-grams. What is the running-time complexity of these algorithms?