Mehryar Mohri
Foundations of Machine Learning 2024
Courant Institute of Mathematical Sciences
Homework assignment 3
November 5, 2024
Due: November 19, 2024

A Kernel methods

1. For $\alpha \geq 0$, the kernel $K_\alpha\colon (x, x') \mapsto \sum_{k=1}^N \min(|x_k|^\alpha, |x'_k|^\alpha)$ over $\mathbb{R}^N \times \mathbb{R}^N$ is used in image classification. Show that K_α is PDS. To do that, you can proceed as follows.

   (a) Use the fact that $(f, g) \mapsto \int_{t=0}^{+\infty} f(t)g(t)dt$ is an inner product over the set of measurable functions over $[0, +\infty)$ to show that $(x, x') \mapsto \min(x, x')$ is a PDS kernel (hint: associate an indicator function to x and another one to x').

   (b) Use the previous question to show that K_1 is PDS and similarly K_α with other values of α.

   Solution:

   (a) Observe that $\min(|u|^\alpha, |u'|^\alpha) = \int_0^{+\infty} 1_{t\in[0,|u'|^\alpha]} 1_{t\in[0,|u'|^\alpha]} dt$, which shows that $(u, u') \mapsto \min(|u|^\alpha, |u'|^\alpha)$ is PDS.

   (b) Since $K_\alpha(x, x') = \sum_{k=1}^N \min(|x_k|^\alpha, |x'_k|^\alpha)$, K_α is PDS as a sum of N PDS kernels.

B Boosting

1. In class, we showed that AdaBoost can be viewed as coordinate descent applied to a convex upper bound on the empirical error. Here, we consider instead an algorithm seeking to minimize the empirical margin loss. For any $0 \leq \rho < 1$, using the same notation as in class, let $\widehat{R}_\rho(f) = \frac{1}{m}\sum_{i=1}^m 1_{y_i f(x_i) \leq \rho}$ denote the empirical margin loss of a function f of the form $f = \frac{\sum_{t=1}^T \alpha_t h_t}{\sum_{t=1}^T \alpha_t}$ for a labeled sample $S = ((x_1, y_1), \ldots, (x_m, y_m))$.

   (a) Prove the upper bound $\widehat{R}_\rho(f) \leq \exp\big(\sum_{t=1}^T \alpha_t \rho\big) \prod_{t=1}^T Z_t$, where the normalization factors Z_t are defined as in the case of AdaBoost in class.

   (b) Give the expression of Z_t as a function of ρ and ϵ_t, where the weighted error ϵ_t are defined as in the case of AdaBoost in class. Use that to prove the following upper bound

   $$\widehat{R}_\rho(f) \leq \exp\bigg(-\sum_{t=1}^T \mathsf{D}\Big(\frac{1-\rho}{2}\Big\|\epsilon_t\Big)\bigg),$$

   where $\mathsf{D}(p\|q)$ denotes the binary relative entropy of p and q: $\mathsf{D}(p\|q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$, for any $p, q \in [0, 1]$.

   (c) Assume that for all $t \in [1, T]$, $\frac{1-\rho}{2} - \epsilon_t > \gamma > 0$. Use the result of the previous question to show that

   $$\widehat{R}_\rho(f) \leq \exp\big(-2\gamma^2 T\big).$$

   (hint: you can use Pinsker's inequality: $\mathsf{D}(p\|q) \geq 2(p-q)^2$ for all $p, q \in [0, 1]$). Show that for $T > \frac{\log m}{2\gamma^2}$, all points have margin at least ρ.

Solution:

(a) First, we show $\widehat{R}_\rho(f)$ can be upper-bounded as follows:

$$\widehat{R}_\rho(f) = \frac{1}{m}\sum_{i=1}^m 1_{y_i f(x_i)\le\rho} = \frac{1}{m}\sum_{i=1}^m 1_{y_i \sum_{t=1}^T \alpha_t h_t(x_i) - \rho\sum_{t=1}^T \alpha_t \le 0}$$
$$\le \frac{1}{m}\sum_{i=1}^m \exp\left(-y_i \sum_{t=1}^T \alpha_t h_t(x_i) + \rho \sum_{t=1}^T \alpha_t\right).$$

Let D_1 be the uniform distribution, that is $D_1(i) = \frac{1}{m}$ for all $i \in [1,m]$ and for any $t \in [2,T]$, define D_t by

$$D_t(i) = \frac{D_{t-1}(i)\exp(-y_i\alpha_{t-1}h_{t-1}(x_i))}{Z_{t-1}},$$

with $Z_{t-1} = \sum_{i=1}^m D_{t-1}(i)\exp(-y_i\alpha_{t-1}h_{t-1}(x_i))$. Observe that $D_t(i) = \frac{\exp\left(-y_i\sum_{s=1}^{t-1}\alpha_s h_s(x_i)\right)}{m\prod_{s=1}^{t-1} Z_t}$. Thus, we can write

$$\begin{aligned}
\widehat{R}_\rho(f) &\le \frac{1}{m}\sum_{i=1}^m \exp\left(-y_i \sum_{t=1}^T \alpha_t h_t(x_i) + \rho \sum_{t=1}^T \alpha_t\right) \\
&= \frac{1}{m}\sum_{i=1}^m \left(m\prod_{t=1}^T Z_t\right) D_t(i) \exp\left(\rho\sum_{t=1}^T \alpha_t\right) \\
&= \exp\left(\rho\sum_{t=1}^T \alpha_t\right)\left(\prod_{t=1}^T Z_t\right).
\end{aligned}$$

(b) The normalization factor Z_t can be expressed in terms of ϵ_t and ρ using its definition:

$$\begin{aligned}
Z_t &= \sum_{i=1}^m D_t(i)\exp(-y_i\alpha_t h_t(x_i)) \\
&= e^{-\alpha_t}(1-\epsilon_t) + e^{\alpha_t}\epsilon_t \\
&= \sqrt{\frac{1+\rho}{1-\rho}(1-\epsilon_t)\epsilon_t} + \sqrt{\frac{1-\rho}{1+\rho}(1-\epsilon_t)\epsilon_t} \\
&= \sqrt{\epsilon_t(1-\epsilon_t)}\left[\sqrt{\frac{1+\rho}{1-\rho}} + \sqrt{\frac{1-\rho}{1+\rho}}\right] \\
&= \sqrt{\epsilon_t(1-\epsilon_t)}\left[\frac{2}{\sqrt{1-\rho^2}}\right] \\
&= 2\sqrt{\frac{\epsilon_t(1-\epsilon_t)}{1-\rho^2}}.
\end{aligned}$$

Define u by $u = \frac{1-\rho}{1+\rho}$. Plugging in that expression in the bound of the previous question and using the expression of α_t gives

$$\begin{aligned}
\widehat{R}_\rho(f) &\le (\prod_t e^{\alpha_t})^\rho \left(\prod_{t=1}^T \sqrt{\epsilon_t(1-\epsilon_t)}(u^{\frac{1}{2}} + u^{-\frac{1}{2}})\right) \\
&= \left(\sqrt{\frac{1-\rho}{1+\rho}}\right)^{\rho T}\left(\prod_t \sqrt{\frac{1-\epsilon_t}{\epsilon_t}}\right)^\rho \left(\prod_{t=1}^T \sqrt{\epsilon_t(1-\epsilon_t)}(u^{\frac{1}{2}} + u^{-\frac{1}{2}})\right) \\
&= \left(u^{\frac{1+\rho}{2}} + u^{-\frac{1-\rho}{2}}\right)^T \prod_{t=1}^T \sqrt{\epsilon_t^{1-\rho}(1-\epsilon_t)^{1+\rho}}.
\end{aligned}$$

Observe that

$$\begin{aligned} u^{\frac{1+\rho}{2}} + u^{-\frac{1-\rho}{2}} &= \Big(\frac{1-\rho}{1+\rho}\Big)^{\frac{1+\rho}{2}} + \Big(\frac{1+\rho}{1-\rho}\Big)^{\frac{1-\rho}{2}} \\ &= \frac{(1-\rho)+(1+\rho)}{(1+\rho)^{\frac{1+\rho}{2}}(1-\rho)^{\frac{1-\rho}{2}}} \\ &= \frac{2}{(1+\rho)^{\frac{1+\rho}{2}}(1-\rho)^{\frac{1-\rho}{2}}} \\ &= \frac{1}{\big(\frac{1+\rho}{2}\big)^{\frac{1+\rho}{2}}\big(\frac{1-\rho}{2}\big)^{\frac{1-\rho}{2}}}. \end{aligned}$$

We also have

$$\begin{aligned} &\log\Big[\sqrt{\epsilon_t^{1-\rho}(1-\epsilon_t)^{1+\rho}}\Big] \\ &= \frac{1-\rho}{2}\log(\epsilon_t) + \frac{1+\rho}{2}\log(1-\epsilon_t) \\ &= -\mathsf{D}\Big(\frac{1-\rho}{2}\Big\|\epsilon_t\Big) + \frac{1-\rho}{2}\log\Big(\frac{1-\rho}{2}\Big) + \frac{1+\rho}{2}\log\Big(\frac{1+\rho}{2}\Big) \\ &= -\mathsf{D}\Big(\frac{1-\rho}{2}\Big\|\epsilon_t\Big) + \log\Big(\Big(\frac{1+\rho}{2}\Big)^{\frac{1+\rho}{2}}\Big(\frac{1-\rho}{2}\Big)^{\frac{1-\rho}{2}}\Big). \end{aligned}$$

Combining these two inequalities gives

$$\widehat{R}_\rho(f) \le \exp\Big(-\sum_{t=1}^{T}\mathsf{D}\Big(\frac{1-\rho}{2}\Big\|\epsilon_t\Big)\Big).$$

(c) By Pinsker's inequality, we have $\mathsf{D}\big(\frac{1-\rho}{2}\big\|\epsilon_t\big) \ge 2\big[\frac{1-\rho}{2} - \epsilon_t\big]^2$. Thus, we can write

$$\widehat{R}_\rho(f) \le \exp\big(-2\gamma^2 T\big).$$

Thus, if the upper bound is less that $1/m$, then $\widehat{R}_\rho(f) = 0$ and every training point has margin at least ρ. The inequality $\exp\big(-2\gamma^2 T\big) < 1/m$ is equivalent to $T > \frac{\log m}{2\gamma^2}$.

C Maxent

1. Derive optimization problem of L_2-regularized Maxent with Mahalonobis distance (counterpart of Maxent with relative entropy). Show that it is a convex optimization problem.
2. Give the general form of the solution (it might be useful to use Lagrange function and representer theorem).
3. Derive dual problem and equivalence (Lagrange duality).

Solution:

1. The optimization problem can be expressed as

$$\min_{\mathbf{p}\in\Delta} (\mathbf{p}-\mathbf{p}_0)^\top \mathbf{K}^{-1}(\mathbf{p}-\mathbf{p}_0)$$

$$\text{subject to: } \Big\| \mathop{\mathbb{E}}_{x\sim\mathbf{p}}[\mathbf{\Phi}(x)] - \mathop{\mathbb{E}}_{x\sim\hat{\mathcal{D}}}[\mathbf{\Phi}(x)]\Big\|_2^2 \le \lambda.$$

which is a convex optimization problem by following the optimization slides taught in class.

2. You can refer to the L_2-squared regularized maxent and the corresponding derivation taught in class.
3. You can directly solve it using the standard method of Lagrange multipliers taught in class.