NYU CSCI-GA.2566-001: Foundations of Machine Learning Fall 2020

Homework 3

Name: R. Teal Witter

A. Kernel Methods

1. Let $W = (W_{p,q})$ be a symmetric $n \times n$ matrix where $W_{p,q}$ is the weight of the edge between nodes p and q and n is the number of nodes in the graph. Observe that $W_{p,q}$ if there is no edge between p and q. Now consider the matrix $K = W^TW$. An entry $K(p,q)$ is
$$\sum_{u \in V} W^T_{p,u} W_{u,q}. \tag{1}$$
Observe that Equation 1 is exactly the sum of the weights of all paths of length two between p and q: If u is not on a path of length two between p and q, then either u does not share an edge with p or with q and so u contributes 0 to the sum. If u is on a path of length two between p and q, then there is an edge between p and u and u and q and so u contributes the product of the weights to the sum.

   Now we want to show that K is PDS. Let $\mathbf{c} \in \mathbb{R}^{n \times 1}$. Then
$$\mathbf{c}^T W^T W \mathbf{c} = ||W\mathbf{c}||_2^2 \geq 0.$$
Therefore K is a PDS kernel.

2. Pixel kernel

   (a) Note that
$$\int_{t=0}^{\infty} 1_{t \in [0,z]} 1_{t \in [0,z']} dt = \langle 1_{[0,z]}, 1_{[0,z']} \rangle.$$
When z or z' is negative, then S returns 0 since $t \notin [0,z], [0,z']$ when $t \geq 0$. Otherwise, $S(z,z') = \min\{z,z'\}$ since we only have non-zero values when $0 \leq t \leq z, z'$.

   To show that S is PDS, let $\mathbf{c}$ be any real vector. Then
$$\langle \mathbf{c}1_{[0,z]}, \mathbf{c}1_{[0,z']} \rangle = ||c||_2^2 \langle 1_{[0,z]}, 1_{[0,z']} \rangle \geq 0$$
since every element in $1_{[0,z]}$ and $1_{[0,z']}$ is nonnegative. Therefore S is a PDS kernel.

   (b) Fix $k \in [N]$. From (a), we know that $\min\{|x_k|^\mu, |x'_k|^\mu\} = S(|x_k|^\mu, |x'_k|^\mu)$. (Since $|x_k|^\mu$ and $|x'_k|^\mu$ are both positive, we don't have to worry about negative values.)

   Then $\exp(\min\{|x_k|^\mu, |x'_k|^\mu\})$ is also PDS from Theorem 6.10 (PDS kernels - closure properties) in the book. That is, composition with a power series with non-negative coefficients (i.e. exp) preserves the PDS property.

The final step is to take the product of $\exp(\min\{|x_k|^\mu, |x'_k|^\mu\})$ over $k \in [N]$. The product of two PDS kernels is also PDS (again by Theorem 6.10). By repeated application of this property, we can extend it to the product of any finite number of PDS kernels. Therefore

$$\prod_{k=1}^{N} e^{\min\{|x_k|^\mu, |x'_k|^\mu\}}$$

is PDS.

B. Boosting

1. Logistic loss boosting

   (a) To show that Φ is convex we show that its second derivative is positive for $u \in \mathrm{dom}(\Phi)$ and $\mathrm{dom}(\Phi)$ is a convex set. The domain of Φ is the set of real numbers $\mathbb{R}$. The set of real numbers is clearly convex: for $x, y \in \mathbb{R}$, any number between x and y must also be in $\mathbb{R}$. We take the derivatives of Φ:

$$\begin{aligned}\Phi(u) &= \log_2(1+e^{-u})\\ \Phi'(u) &= \frac{1}{\ln(2)}\frac{-e^{-u}}{1+e^{-u}}\\ \Phi''(u) &= -\frac{1}{\ln(2)}\frac{-(1+e^{-u})e^{-u} - e^{-u}e^{-u}}{(1+e^{-u})^2} = \frac{1}{\ln(2)}\frac{1}{(1+e^{u})^2} \geq 0\end{aligned}$$

for all $u \in \mathbb{R}$. Therefore Φ is convex.
To see that Φ is decreasing, take $x < y$ for $x, y \in \mathbb{R}$. Then

$$\begin{aligned}-x > -y &\Leftrightarrow e^{-x} > e^{-y}\\ &\Leftrightarrow 1+e^{-x} > 1+e^{-y}\\ &\Leftrightarrow \log_2(1+e^{-x}) > \log_2(1+e^{-y})\\ &\Leftrightarrow \Phi(x) > \Phi(y)\end{aligned}$$

since both $\log_2$ and exp are monotone increasing. Therfore Φ is decreasing.
To see that Φ upper bounds the zero-one loss, take $0 \geq x$. Then

$$\begin{aligned}0 \leq -x &\Leftrightarrow e^0 = 1 \leq e^{-x}\\ &\Leftrightarrow \log_2(1+1) \leq \log_2(1+e^{-x})\\ &\Leftrightarrow 1 \leq \Phi(x).\end{aligned}$$

Now take $0 \leq x$,

$$\begin{aligned}e^{-x} \geq 0 &\Leftrightarrow 1+e^{-x} \geq 1\\ &\Leftrightarrow \log_2(1+e^{-x}) \geq 0\\ &\Leftrightarrow \Phi(x) \geq 0\end{aligned}$$

where the the first inequality follows from the positivity of e^y for all $y \in \mathbb{R}$. Together, $\Phi(x) \geq 1$ when x is negative (i.e. the zero-one loss "fires") and $\Phi(x) \geq 0$ when x is positive (i.e. the zero-one loss stays at 0). Thefore Φ upperbounds the zero-one loss.

(b) Let the function $f(x) = \sum_{j=1}^{N} \alpha_j h_j(x)$ for a given N-length vector of non-negative coefficients α where h_j is in the hypothesis set H and H has cardinality N. For pairs of points and labels (x_i, y_i) where $i \in [m]$, define the objective function

$$F(\alpha) = \frac{1}{m} \sum_{i=1}^{m} \log_2(1 + e^{-y_i \sum_{j=1}^{p} \alpha_j h_j(x_i)}).$$

We now argue that F is convex with respect to α: $-y_i f(x_i)$ is convex because it is an affine function of α. Φ is convex by (a) and so $\log_2(1+\exp(-y_i f(x_i)))$ is also convex since composition with a monotone increasing convex function (i.e. Φ) preserves convexity. Finally the sum of convex functions is convex and multiplying by a scalar does not affect convexity. Therefore F is convex with respect to α.

(c) For $t \in [T]$ and $k \in [N]$, define

$$\begin{aligned} f_t &= \sum_{j=1}^{N} \alpha_t h_j(x_i) \\ Z_t &= \sum_{i=1}^{m} \frac{e^{-y_i f_{t-1}(x_i)}}{\ln(2)(1 + e^{-y_i f_{t-1}(x_i)})} \\ D_t(i) &= \frac{e^{-y_i f_{t-1}(x_i)}}{\ln(2)(1 + e^{-y_i f_{t-1}(x_i)}) Z_t} \\ \epsilon_{t,k} &= \mathop{\mathbb{E}}_{i \sim D_t} [1_{y_i h_k(x_i)} \leq 0]. \end{aligned}$$

The directional derivative of F at α_{t-1} along e_k is defined by

$$\begin{aligned} F'(\alpha_{t-1}, e_k) &= \lim_{\eta \to 0} \frac{F(\alpha_{t-1}, e_k) - F(\alpha_{t-1})}{\eta} \text{ where} \\ F(\alpha_{t-1}, \eta e_k) &= \frac{1}{m} \sum_{i=1}^{m} \log_2(1 + e^{-y_i \sum_{j=1}^{N} \alpha_{t-1} h_j(x_i) - \eta y_i h_k(x_i)}). \end{aligned}$$

Then taking the derivative with respect to η and immediately setting η to 0 yields

$$\begin{aligned} F'(\alpha_{t-1}, e_k) &= -\frac{1}{m} \sum_{i=1}^{m} \frac{y_i h_k(x_i) e^{-y_i f_{t-1}(x_i)}}{\ln(2)(1 + e^{-y_i f_{t-1}(x_i)})} \\ &= -\frac{1}{m} \sum_{i=1}^{m} y_i h_k(x_i) D_t Z_t \\ &= -\frac{Z_t}{m} \left[\sum_{i=1}^{m} D_t(i) 1_{y_i h_k(x_i) = 1} - \sum_{i=1}^{m} D_t(i) 1_{y_i h_k(x_i) = -1} \right] \\ &= -\frac{Z_t}{m} \left[(1 - \epsilon_{e,k}) - \epsilon_{e_k} \right] = \frac{Z_t}{m} \left[2\epsilon_{t,k} - 1 \right]. \end{aligned} \tag{2}$$

We want to maximize the absolute value of Equation 2 to get the best descent so we want the smallest value of $\epsilon_{t,k}$ (since Z_t and m do not change with respect to the choice of k). Therefore when boosting on the tth iteration we want to find k such that $\epsilon_{t,k}$ is smallest. Observe that our weak learning condition must be that there is some k such that $\epsilon_{t,k} \leq 1/2$ on the tth iteration, for the distribution D_t, and for the sample of m points.

(d) Fix $(u, v) \in \mathbb{R}^2$, then

$$\begin{aligned}
\Phi(u+v) - \Phi(u) &= \log_2(1 + e^{-u-v}) - \log_2(1 + e^{-u}) \\
&= \log_2\left(\frac{1 + e^{-u-v}}{1 + e^{-u}}\right) \\
&= \log_2\left(\frac{(e^{-v} - 1)e^{-u} + 1 + e^{-u}}{1 + e^{-u}}\right) \\
&= \log_2\left(\frac{(e^{-v} - 1)e^{-u}}{1 + e^{-u}} + 1\right). \qquad (3)
\end{aligned}$$

Now define $x = (e^{-v} - 1)e^{-u}/(1 + e^{-u})$. Observe that $x \geq -1$ since $e^{-v} \geq 0$ and $1 \geq e^{-u}/(1 + e^{-u}) \geq 0$ for all $u, v \in \mathbb{R}$.

We want to show that $f(x) = x - \ln(x + 1) \geq 0$ for $x \geq -1$. Together

$$f'(x) = 1 - 1/(x + 1) = 0 \Leftrightarrow x = 0$$

and

$$f''(x) = 1/(x + 1)^2 > 0$$

imply that $x = 0$ is minimal point. Since

$$\lim_{x \to -1^+} f(x) = \infty = \lim_{x \to \infty} f(x)$$

$x = 0$ is indeed a global minimum. Then $f(0) = 0$ implies that $f(x) \geq 0$ so

$$\ln(x + 1) \leq x \qquad (4)$$

for $x \geq -1$.

Continuing from Equation 3, we use Equation 4 and get that

$$\begin{aligned}
\Phi(u+v) - \Phi(u) &= \ln\left(\frac{(e^{-v} - 1)e^{-u}}{1 + e^{-u}} + 1\right) \\
&\leq \ln\left(\frac{(e^{-v} - 1)e^{-u}}{1 + e^{-u}} + 1\right)\frac{1}{\ln(2)} \\
&\leq \frac{(e^{-v} - 1)e^{-u}}{\ln(2)(1 + e^{-u})} \\
&= -\frac{-e^{-u}}{\ln(2)(1 + e^{-u})}(e^{-v} - 1) = -\Phi'(u)(e^{-v} - 1).
\end{aligned}$$

Therefore

$$\Phi(u+v) - \Phi(u) \leq -\Phi'(u)(e^{-v} - 1) \qquad (5)$$

for all $(u, v) \in \mathbb{R}^2$.

(e) We have

$$\begin{aligned}
F(\alpha_{t-1} + \eta e_k) - F(\alpha_{t-1}) = \frac{1}{m}\sum_{i=1}^{m} \log_2(1 + e^{-y_i \sum_{j=1}^{N} \alpha_{t-1,j} h_j(x_i) - \eta y_i h_k(x_i)}) \\
- \frac{1}{m}\sum_{i=1}^{m} \log_2(1 + e^{-y_i \sum_{j=1}^{N} \alpha_{t-1,j} h_j(x_i)}).
\end{aligned}$$

Define $u_i = y_i \sum_{j=1}^{N} \alpha_{t-1,j} h_j(x_i)$ and $v_i = \eta y_i h_k(x_i)$. Then using Equation 5

$$\begin{aligned}
F(\alpha_{t-1} + \eta e_k) - F(\alpha_{t-1}) &= \frac{1}{m} \sum_{i=1}^{m} \Phi(u_i + v_i) - \Phi(u_i) \\
&\leq \frac{1}{m} \sum_{i=1}^{m} -\Phi'(u_i)(e^{-v_i} - 1) \\
&= \frac{1}{m} \sum_{i=1}^{m} -\frac{-e^{-u_i}}{\ln(2)(1 + e^{u_i})}(e^{-v_i} - 1) \\
&= \frac{1}{m} \sum_{i=1}^{m} D_t(i) Z_t (e^{-\eta y_i h_k(x_i)} - 1).
\end{aligned}$$

Therefore

$$F(\alpha_{t-1} + \eta e_k) - F(\alpha_{t-1}) \leq \frac{1}{m} \sum_{i=1}^{m} D_t(i) Z_t (e^{-\eta y_i h_k(x_i)} - 1). \tag{6}$$

(f) We minimize the upper bound in Equation 6 by differentiating with respect to η and setting the result equal to 0. We get

$$\begin{aligned}
\frac{1}{m} \sum_{i=1}^{m} D_t(i) Z_t (-y_i h_k(x_i) e^{-\eta y_i h_k(x_i)}) &= 0 \\
-\frac{Z_t}{m} \left[\sum_{i=1}^{m} D_t(i) 1_{y_i h_k(x_i)=1} e^{-\eta} - \sum_{i=1}^{m} D_t(i) 1_{y_i h_k(x_i)=-1} e^{\eta} \right] &= 0 \\
(1 - \epsilon_{t,k}) e^{-\eta} - \epsilon_{t,k} e^{\eta} &= 0 \\
\frac{1 - \epsilon_{t,k}}{\epsilon_{t,k}} &= e^{2\eta} \\
\frac{1}{2} \ln \left(\frac{1 - \epsilon_{t,k}}{\epsilon_{t,k}} \right) &= \eta. \qquad (7)
\end{aligned}$$

Equation 7 is exactly the same as the minimization step size (there the variable is α_t) in AdaBoost since our $\epsilon_{t,k}$ is their ϵ_t. (Their ϵ_t is defined to be in the best direction k.) At iteration t, the step is given by $\alpha_t = \alpha_{t-1} + \eta e_k$ where η is the step size given in Equation 7 and e_k is the step direction chosen because its error is smallest. In terms of $f_t = \sum_{i=j}^{N} \alpha_{t,j} h_j$, the step is given by $f_t = f_{t-1} + \eta h_k = f_{t-1} + \alpha_t h_t$ where $\alpha_t = \eta$ and $h_t = h_k$ are alternate notation for the step size and direction.

(g) The pseudocode appears in Algorithm 1.

Algorithm 1: Logistic loss boosting.

```
input: set of samples $S = ((x_1, y_1), \ldots, (x_m, y_m))$,
       number of iterations $T$,
       hypothesis set $H$ with $N$ base predictors
output: function $f$
$f_0 \leftarrow 0$
for $i \leftarrow 1$ to $m$ do
    $D_1(i) \leftarrow 1/m$
for $t \leftarrow 1$ to $T$ do
    $h_t \leftarrow \arg\min_{h \in H} \Pr_{i \sim D_t}(h(x_i) \neq y_i)$
    $\epsilon_t \leftarrow \Pr_{i \sim D_t}(h_t(x_i) \neq y_i)$
    $\alpha_t \leftarrow \frac{1}{2} \ln(\frac{1-\epsilon_t}{\epsilon_t})$
    $f_t \leftarrow f_{t-1} + \alpha_t h_t$
    $Z_{t+1} \leftarrow \sum_{i=1}^{m} e^{-y_i f_t(x_i)}/(\ln(2)(1 + e^{-y_i f_t(x_i)}))$
    for $i \leftarrow 1$ to $m$ do
        $D_{t+1}(i) \leftarrow e^{-y_i f_t(x_i)}/(\ln(2)(1 + e^{-y_i f_t(x_i)})Z_{t+1})$
return $f_T$
```

(h) We will use Corollary 7.5 and Corollary 7.6 from the book. They cannot be directly applied to the function returned from Algorithm 1 since it is not a convex combination of base hypotheses. Instead, define $\bar{f} = \sum_{t=1}^{T} \alpha_t h_t / (||\alpha||_1) \in \text{conv}(H)$ where α is the vector with 0 entries in the positions corresponding to h that are not picked in any iteration $t \in [T]$ and α_t in the positions corresponding to h_t that are picked at iteration t. Observe that $\bar{f}$ and f are equivalent in the context of binary classification since $\text{sgn}(f) = \text{sgn}(f/||\alpha||_1)$. It follows that $R(f) = R(f/||\alpha||_1)$.

Then for H a set of real-valued functions, margin $\rho > 0$, any $\delta > 0$, and $\bar{f} \in \text{conv}(H)$, Corollary 7.5 (Ensemble Rademacher margin bound) yields

$$R(f) = R(\bar{f}) \leq \hat{R}_{S,\rho}(\bar{f}) + \frac{2}{\rho}\mathcal{R}_m(H) + \sqrt{\frac{\ln(1/\delta)}{2m}}$$

with probability $1 - \delta$.

We also have the option to bound the Rademacher complexity like so

$$\mathcal{R}_m(H) \leq \sqrt{\frac{2d \log(em/d)}{m}}$$

as in Corollary 7.6 (Ensemble VC-Dimension margin bound).

Define

$$\rho_f = \min_{i \in [m]} \frac{|\alpha \cdot f(x_i)|}{||\alpha||_1}.$$

Then, since the margin loss can be upper bounded by the fraction of points x labeled with y in the training sample with confidence margin at most ρ, we can write

$$\hat{R}_{S,\rho}(\bar{f}) \leq \frac{|\{i \in [m] : y_i \rho_f(x_i) \leq \rho\}|}{m}.$$

Since $\Phi : u \to \log_2(1 + e^{-u})$ upper bounds the zero-one loss by part (a), we can also write

$$\begin{aligned}\hat{R}_{S,\rho}(\bar{f}) &= \frac{1}{m}\sum_{i=1}^{m} 1_{y_i f(x_i) - \rho||\alpha||_1 \leq 0} \\ &\leq \frac{1}{m}\sum_{i=1}^{m} \log_2(1 + e^{-y_i f(x_i) + \rho||\alpha||_1}).\end{aligned}$$

(i)