Introduction to Machine Learning (67577)
Exercise VI – PCA, kernels, SGD and DL
Hadar Sharvit – 208287599

Theoretical Questions

PCA

1. Let us first calculate $Var[\langle v, X\rangle] = E[(\langle v, X\rangle - E[\langle v, X\rangle])^2]$. Firstly, lets write $X = (x_1, \dots, x_d)^T$, where x_i are sampled from some distribution over $\mathbb{R}^d$. If so, we can write
$$E[\langle v, X\rangle] = E[v_1 x_1 + \dots + v_d x_d] = \sum_{i=1}^{d} E[v_i x_i] = \sum_{i=1}^{d} v_i E[x_i] = E[X]\sum_{i=1}^{d} v_i = 0 \cdot \sum v_i = 0$$
we are left with $E[\langle v, X\rangle^2] = E[\langle v, X\rangle\langle X, v\rangle] = v^T E[X \cdot X^T]v = v^T \Sigma v$. From here we can say that given a leading eigenvector of Σ, v' $v^T\Sigma v \le v'^{\mathrm{T}}\Sigma v'$. In other words, the variance of $\langle v, X\rangle$ is $\le$ the variance of the PCA embedding of X into $1 - \dim$

Kernels

2. Given valid kernel $k(x, x')$, provide a normalized kernel $\tilde{k}$ such that for all x, $\tilde{k}(x, x) = 1$:
We define
$$\tilde{k}(x, x') = \frac{k(x, x')}{\sqrt{k(x, x) \cdot k(x', x')}}$$
Firstly, $\tilde{k}$ is a valid kernel, as it represents an inner product of the feature map $\psi/||\psi||$:
$$\tilde{k}(x, x') = \frac{\langle\psi(x), \psi(x')\rangle}{\sqrt{\langle\psi(x), \psi(x)\rangle\langle\psi(x'), \psi(x')\rangle}} = \frac{\langle\psi(x), \psi(x')\rangle}{\sqrt{||\psi(x)||^2||\psi(x')||^2}} = \frac{\langle\psi(x), \psi(x')\rangle}{||\psi(x)|| \cdot ||\psi(x')||}$$
$$= \left\langle \frac{\psi(x)}{||\psi(x)||}, \frac{\psi(x')}{||\psi(x')||} \right\rangle$$
Furthermore, $\tilde{k}(x, x) = 1$ for every x, as
$$\tilde{k}(x, x) = \frac{\langle\psi(x), \psi(x)\rangle}{||\psi(x)|| \cdot ||\psi(x)||} = \frac{||\psi(x)||^2}{||\psi(x)||^2} = 1$$

3. Example of a data set S and a feature map $\psi: \mathbb{R}^d \to \mathcal{F}$ where S is not linearly separable in $\mathbb{R}^{d\ge 2}$ but the transformed data $\psi(s)$ $\forall s \in S$ is linearly separable in $\mathcal{F}$:
Consider $d = 2$, $S = S_1 \cup S_2$ where $S_i = \left\{(x, y): r_i \le \sqrt{x^2 + y^2} \le r_{i+1}\right\}$ represents a set of points within some ring with inner radius r_i and outer radius r_{i+1}, where $r_1 < r_2 < r_3$. We define the mapping as followed: given some point $(x, y) \in S$, $\psi(x, y) = \left(x, y, \sqrt{x^2 + y^2}\right)$. Simply put, ψ maps the point (x, y) to the same (x, y) coordinate, with some elevation $z = \sqrt{x^2 + y^2}$ to it. Since $r_1 < r_2 < r_3$, such mapping provides a set of linearly separable data points in $\mathbb{R}^3$

Convex optimization

4.
a. Show that if $f_i: V \to \mathbb{R}$ are convex ($\forall i \in [m]$), then $g(u) = \sum_{i=1}^{m} \gamma_i f_i(u)$, ($\gamma_i \ge 0$) is convex:

Since f_i is convex, we know that $\forall v_1, v_2 \in V$, $\forall \alpha \in [0,1]$ $\alpha f_i(v_1) + (1-\alpha) f(v_2) \geq f_i(\alpha v_1 + (1-\alpha) v_2)$. If so, $\alpha \gamma_i f_i(v_1) + (1-\alpha)\gamma_i f(v_2) \geq \gamma_i f_i(\alpha v_1 + (1-\alpha) v_2)$. In other words, $\gamma_i f_i$ is also convex $\forall i \in [m]$ by definition. If so, we have

$$\alpha g(v_1) + (1-\alpha) g(v_2) = \alpha \sum_{i=1}^{m} \gamma_i f_i(v_1) + (1-\alpha) \sum_{i=1}^{m} \gamma_i f_i(v_2)$$
$$= \sum_{i=1}^{m} \alpha \gamma_i f_i(v_1) + (1-\alpha)\gamma_i f_i(v_2) \geq \sum_{i=1}^{m} \gamma_i f_i(\alpha v_1 + (1-\alpha) v_2)$$
$$= g(\alpha v_1 + (1-\alpha) v_2)$$

b. Composition of convex function is not necessarily convex:
Take $f(x) = -x$, $g(x) = e^x \to h(x) = f\big(g(x)\big) = -e^x$. Both f and g are convex, yet h is not.

c. Given convex set C, a function $f: C \to \mathbb{R}$ is convex $\Leftrightarrow$ $\text{epigraph}(f) = \{(u,t): f(u) \leq t\}$ is a convex set:
In the first direction, assume f is convex. If so, $\forall c_1, c_2 \in C$, $\forall \alpha \in [0,1]$ we have $\alpha f(c_1) + (1-\alpha) f(c_2) \geq f(\alpha c_1 + (1-\alpha) c_2)$. Assume towards contradiction that $\text{epi}(f)$ is not convex, therefore $\exists (x,y), (z,t) \in \text{epi}(f)$ $\exists \alpha \in [0,1]$ such that $\alpha(x,y) + (1-\alpha)(z,t) = (\alpha x, \alpha y) + \big((1-\alpha)z, (1-\alpha)t\big) = (\alpha x + (1-\alpha)z, \alpha y + (1-\alpha)t) \notin \text{epi}(f)$. In other words, $f(x) \leq y$ and $f(z) \leq t$, yet $f(\alpha x + (1-\alpha)z) > \alpha y + (1-\alpha)t$. On the other hand, since f is convex, then by definition $f(\alpha x + (1-\alpha)z) \leq \alpha f(x) + (1-\alpha) f(z) \leq \alpha y + (1-\alpha)t$. we have reached a contradiction, therefore $\text{epi}(f)$ must be convex.
In the other direction, assume $\text{epi}(f)$ is convex, therefore $\forall (x,y), (z,t) \in \text{epi}(f)$ $\forall \alpha \in [0,1]$ we have $\alpha(x,y) + (1-\alpha)(z,t) = (\alpha x + (1-\alpha)z, \alpha y + (1-\alpha)t) \in \text{epi}(f)$, i.e., $f(\alpha x + (1-\alpha)z) \leq \alpha y + (1-\alpha)t$. The claim holds for every $(x,y), (z,t)$, therefore we can choose $y = f(x)$ and $t = f(z)$, and in that case $f(\alpha x + (1-\alpha)z) \leq \alpha f(x) + (1-\alpha) f(z) \to f$ is convex by definition.

d. Let $f_i: V \to \mathbb{R}$, $i \in I$, $f: V \to \mathbb{R}$, $f(u) = \sup_{i \in I} f_i(u)$. Show that if $\forall i \in I$ f_i is convex, then so is f
From $(4.c)$ it is sufficient to show that $\text{epi}(f) = \text{epi}\left(\sup_{i \in I} f_i(u)\right)$ is a convex set.
Consider some pair (x,y). By definition of epi, $(x,y) \in \text{epi}(f) \Leftrightarrow f(x) \leq y$. Furthermore, since f bounds f_i for every $i \in I$, that is, $f > f_i$ for every $i \in I$, we also have $f_i(x) \leq y$. In other words, $(x,y) \in \text{epi}(f) \Leftrightarrow (x,y) \in \text{epi}(f_i)$ for every $i \in I$, which can only happen if $\text{epi}(f)$ is the intersection of all $\text{epi}(f_i)$. Mathematically speaking, $\text{epi}(f) = \bigcap_{i \in I} \text{epi}(f_i)$. From $(4.c)$ we know that f_i is convex$\to$ $\text{epi}(f_i)$ is a convex set. Since an intersection of convex set is convex we conclude that $\text{epi}(f)$ is convex.

5.

a. Given $x \in \mathbb{R}^d$ and $y \in \{\pm 1\}$, show that the hinge loss function defined by $\ell_{x,y}^{hinge}(w,b) = \max\big(0, 1 - y(w^T x + b)\big)$ is convex (in w, b):
we have shown in class that a pointwise max function is convex. This is not enough as we have to show that both 0 and $1 - y(w^T x + b)$ are convex. This is true because they are both

linear functions, where one of them is a composition with an affine function – all of those were shown to be convex in class

b. Deduce some $g \in \partial \ell_{x,y}^{hinge}(w, b)$:
We need to find a sub-gradient of $\ell_{x,y}^{hin}\ \ (w, b)$, that is, some g for which for all (x, y), we have $\max\left(0{,}1 - y(w^T x + b)\right) \geq \max\left(0{,}1 - y_0(w^T x_0 + b)\right) + \underbrace{\langle g, (x, y) - (x_0, y_0)\rangle}_{=\langle g,(x-x_0,y-y_0)\rangle}$. we can see that $g = \begin{Bmatrix} 0, & \ell_{x_0,y_0}^{hing}\ (w, b) = 0 \\ (-yx, -y), & \ell_{x_0,y_0}^{hinge}(w, b) \neq 0 \end{Bmatrix}$ satisfies the claim.

c. Given $f_1, \dots, f_m: \mathbb{R}^d \to \mathbb{R}$ convex functions, and $\xi_k \in \partial f_k(x)$ for all k. define $f: \mathbb{R}^d \to \mathbb{R}$ by $f(x) = \sum_{i=1}^{m} f_i(x)$. Show that $\sum_k \xi_k \in \partial \sum_k f_k(x)$:
Since ξ_k is a sub-gradient of f_k, we know that $\forall x' \in \mathbb{R}^d\ \ f_k(x') \geq f_k(x) + \langle \xi_k, x' - x\rangle$, therefore the sum also satisfies $\sum_{k=1}^{m} f_k(x') \geq \sum_{k=1}^{m}(f_k(x) + \langle \xi_k, x' - x\rangle)$. Therefore, by the same logic we have by definition that $\sum_k \xi_k \in \partial \sum_k f_k(x)$

d. Given $S = \{(x_i, y_i)\}_{i=1}^{m} \subset \mathbb{R}^d \times \{\pm 1\}$, define $f(w, b) = \frac{1}{m}\sum_{i=1}^{m} \ell_{x_i,y_i}^{hing}\ (w, b) + \frac{\lambda}{2}||w||^2$. Find a member of ∂f for each w:
Same as in $(4.b)$, we can define g_i for every $i \in [m]$ very similar:

$$g_i = \begin{Bmatrix} 0, & \ell_{x_i,y_i}^{hinge}(w, b) = 0 \\ (-y_i x_i, -y_i), & \ell_{x_i,y_i}^{hinge}(w, b) \neq 0 \end{Bmatrix}$$

And using $(4.c)$ we know that

$$\frac{1}{m}\sum_{i=1}^{m} g_i + \lambda(w, 0) \in \partial f$$

<u>Practical Questions</u>