CS446: Machine Learning, Fall 2018, Homework 0

Name: Aahan Agrawal (agrawl10)

Collaborated with Some Person (sperson2), Another Person (aperson3)

Problem (1)

(a)

Note that both H_1 and H_{-1} are parallel hyperplanes, since their normal vectors are equal. It follows that to find the distance between both planes, given a point x_1 lying on H_1, the closest point on H_{-1} to x_1 is the point x_2 obtained from intersecting the line $\{x_1 + wt | t \in \mathbb{R}\}$ with H_{-1}. That is $x_2 = x_1 + wt$. We seek to find $\|wt\|$. Note that $w^T x_1 = 1$ and $w^T x_2 = -1$. Hence:

$$\begin{aligned} x_2 - x_1 &= wt \\ \implies w^T(x_2 - x_1) &= (w^t w)t \\ \implies -2 &= \|w\|^2 t \\ \implies t &= \frac{-2}{\|w\|^2} \\ \implies \|wt\| &= \left\| \left[w \left(\frac{-2}{\|w\|^2} \right) \right] \right\| \\ &= \frac{2}{\|w\|} \end{aligned}$$

(b)

We prove this by contradiction. Suppose that there is a better maximum margin classifier $w' \neq w$ for the set (X^*, Y^*). We will show that w' is also the maximum margin classifier for (X, Y) then, which is a contradiction, since we assumed that w was the maximum margin classifier and the maximum margin classifier is unique [1]

By assumption, since w' is a better classifier for (X^*, Y^*),

$$y^i(w')Tx^i \geq 1 \qquad \text{for all } i \in \mathcal{N} \text{ and} \tag{γ}$$

$$\frac{2}{\|w'\|} \geq \frac{2}{\|w\|}$$

Thus, to prove that w' is the maximum margin classifier for (X, Y), we need to show that

[1] Solving for the maximum margin classifier in the separable case is a strictly convex problem and, hence, the minimizer to the problem is unique.

$$y^j(w')^T x^j \geq 1 \qquad \text{For all } j \notin \mathcal{N}$$

Let $H_1 = \{(x, 1) \in (\mathbb{R}^n, \mathbb{R}^n) | (1)w^T x = 1\}$. Without loss of generality, consider any (x^j, y^j) for $j \notin \mathcal{N}$ such that $y^j = 1$. There exists some $(z, 1) \in H_1$ such that $z + \alpha w = x^j$, where $\alpha > 0$. Thus

$$\begin{aligned} y^j(w')^T x^j &= y^j\alpha(w')^T(z + w) \\ &= y^j(w')^T z + y^j\alpha(w')^T w \\ &\geq 1 + y^j\alpha(w')^T w \\ &= 1 + \alpha(w')^T w \end{aligned}$$

It suffices to show that $(w')^T w \geq 0$, to prove that (x^j, y^j) is correctly classified using w'.

Observe that for any $\beta > 0$, we must have that βw is classified using the maximum margin classifier under w, which we label $\mathcal{A}_w$, as having label 1. For if q is the label of βw, then

$$\mathcal{A}_w(\beta w) = q(w^T(\beta w)) > 0 \implies q = 1$$

Thus, in particular, there exists some $\beta' > 0$ such that $\mathcal{A}_w(\beta' w) = 1$, meaning that $\beta' w$ lies on H_1 and we must have, by construction, $(w')^T(\beta' w) > 0 \implies (w')^T w \geq 0$.

Thus w' correctly classifies even (x^j, y^j). Since $j \notin \mathcal{N}$ was arbitrary, $\mathcal{A}'_w$ correctly classifies all $j \notin \mathcal{N}$. $\mathcal{A}'_w$ already correctly classfied those $(x^i, y^i) \in \mathcal{N}$ and $\frac{2}{\|w'\|} \geq \frac{2}{\|w\|}$, so we conclude that w' is a better classifier than w for (X, Y), which is a contradiction.

CS446: Machine Learning, Fall 2018, Homework 2

Name: Aahan Agrawal (agrawl10)

Collaborated with Some Person (sperson2), Another Person (aperson3)

Problem (2)

(a)

Since A is symmetric, A is unitarily diagonalizable, meaning that

$$A = P^T D P$$

Where $P^T = P^{-1}$ and D is diagonal

Since D is diagonal and A positive semi-definite, the eigenvalues of A occupy the diagonal entries of D and all eigenvalues are non-negative. As a consequence, we can define a square root for D, which is the matrix obtained by taking the square root of each diagonal entry in D. We call this matrix E

$$\begin{aligned} A &= P^T E E P \\ \implies x^T A x' &= x^T P^T E E P x' \\ &= (EPx)^T (EPx') \end{aligned}$$

Thus, we see that a feature transformation ϕ exists defined by $\phi(x) = EPx$ such that $x^T A x' = k(x, x') \phi(x)^T \phi(x')$.

(b)

Since k is a valid kernel, $k(x, x')$ can be decomposed into the inner product of some feature transformation ϕ. That is, $k(x, x') = \phi(x)^T \phi(x')$.

Define a new feature transformation $\psi(x) = f(x)\phi(x)$. Then observe that

$$\psi(x)^T \psi(x^*) = f(x)\phi(x)^T \phi(x^*) f(x^*)$$

(c)

We show that $x^T K x \geq 0$ for all $x \in \mathbb{R}^n$. Recall that inner products produce non-negative values in $\mathbb{R}$ and that they are symmetric. Thus K is a symmetric matrix with no negative entries. It follows that

$$
\begin{aligned}
x^T A x &= \sum_{i,j} x_i x_j A_{ij} \\
&= 2 \sum_{i,j>i} x_i x_j A_{ij}
\end{aligned}
$$

Since $x_i x_j A_{ij} = x_j x_i A_{ij}$

Now suppose that arbitrary x is given. x can be decomposed as the sum of two vectors x_1 and x_2 such that every entry in x_1 is non-negative and every entry in x_2 is non-positive. It follows that

$$
\begin{aligned}
x^T A x &= (x_1 + x_2)^T A (x_1 + x_2) \\
&= x_1^T A x_1 + x_1^T A x_2 + x_2^T A x_2 + x_2^T A x_2
\end{aligned}
$$

From (1), we know that:

$$
x_1^T A x_1 = 2 \sum_{i,j>i} x_i^i x_j^i A_{ij}
$$

From how we defined x_1, we conclude that this foregoing expression is non-negative. By similar reasoning, we can conclude that $x_2^T A x_2$ is non-negative

By construction, $x_1^T A x_2$ and $x_2^T A x_1$ are both zero, since wherever x_1 is not zero, x_2 is zero and vice versa. Hence

$$
x_1^T A x_1 + x_1^T A x_2 + x_2^T A x_2 + x_2^T A x_2 \geq 0
$$

This completes the proof and the problem.

CS446: Machine Learning, Fall 2018, Homework 0

Name: Aahan Agrawal (agrawl10)

Collaborated with Some Person (sperson2), Another Person (aperson3)

Problem (4)

(a)

From here onwards, we drop the superscript i appearing in $\mathbf{x}^i, \mathbf{y}^i, \mathbf{z}^i$.

$$\mathbf{z} = \mathbf{W_2}\left(\phi\left(\mathbf{W_1}x + \mathbf{b_1}\right)\right) + \mathbf{b_2}$$

Please note that, as a consequence of multiplying by $\mathbf{W_2}$ and $\mathbf{W_1}$, instead of $\mathbf{W_2}^T$ and $\mathbf{W_1}^T$, the weight connecting the hth node in the hidden layer to the kth node in the output layer is $(\mathbf{W_2})_{kh}$ and the weight connecting the dth node in the input layer to the h node in the hidden layer is $(\mathbf{W_2})_{hd}$.

We set up some preliminary results used in both parts (b) and (c)

Differentiating $\mathrm{Err}(\mathbf{y}, \mathbf{z})$ with respect to some weight w (w is a placeholder for any weight apperaing in either $\mathbf{W_1}$, $\mathbf{W_2}$, $\mathbf{b_1}$ or $\mathbf{b_2}$), we obtain:

$$-\sum_{k=1}^{K} y_k \frac{\partial z_k}{\partial w} + \frac{1}{\sum_{k=1}^{K} \exp\left(z_k\right)} \left(\sum_{k=1}^{K} \exp\left(z_k\right) \frac{\partial z_k}{\partial w}\right) \tag{α}$$

We see that the term $\frac{\partial z_k}{\partial w}$ appears repeatedly. Let us express z_k in terms of $\mathbf{W_1}, \mathbf{W_2}, \mathbf{b_1}, \mathbf{b_2}$ and $\mathbf{x}$:

$$z_k = \sum_{j=1}^{h} (\mathbf{W_2})_{kj} \left[\phi\left(\mathbf{W_1}\mathbf{x} + \mathbf{b_1}\right)\right]_j + \left(\mathbf{b_2}\right)_k$$

$$= \sum_{j=1}^{h} (\mathbf{W_2})_{kj}\ \phi\left(\sum_{m=1}^{d} (\mathbf{W_1})_{jm}\mathbf{x}_m + \left(\mathbf{b_1}\right)_j\right) + \left(\mathbf{b_2}\right)_k$$

<++>

(b)

Taking the derivative of z_k with respect to $(\mathbf{W_2})_{kh}$, we obtain:

$$\frac{\partial z_k}{\partial (\mathbf{W_2})_{kh}} = \frac{\partial}{\partial (\mathbf{W_2})_{kh}} \left[\sum_{j=1}^{h} (\mathbf{W_2})_{kj}\ \phi \left(\sum_{m=1}^{d} (\mathbf{W_1})_{jm} \mathbf{x}_m + (\mathbf{b_1})_j \right) + (\mathbf{b_2})_k \right]$$

The only term in $\sum_{j=1}^{h} (\mathbf{W_2})_{kj}$ that is non-zero after differentiation is the term obtained when $j = h$.

$$= \frac{\partial}{\partial (\mathbf{W_2})_{kh}} \left[(\mathbf{W_2})_{kh}\ \phi \left(\sum_{m=1}^{d} (\mathbf{W_1})_{hm} \mathbf{x}_m + (\mathbf{b_1})_h \right) + (\mathbf{b_2})_k \right]$$

$$\frac{\partial z_k}{\partial (\mathbf{W_2})_{kh}} = \left[\phi \left(\sum_{m=1}^{d} (\mathbf{W_1})_{hm} \mathbf{x}_m + (\mathbf{b_1})_h \right) \right]$$

When we take the derivative of z_k with respect to $(\mathbf{b_2})_k$, we observe that the only term involving $(\mathbf{b_2})_k$ is $(\mathbf{b_2})_k$ itself. Hence:

$$\frac{\partial z_k}{\partial (\mathbf{b_2})_k} = \frac{\partial}{\partial (\mathbf{b_2})_k} \left[\sum_{j=1}^{h} (\mathbf{W_2})_{kj}\ \phi \left(\sum_{m=1}^{d} (\mathbf{W_1})_{jm} \mathbf{x}_m + (\mathbf{b_1})_j \right) + (\mathbf{b_2})_k \right]$$

$$\frac{\partial z_k}{\partial (\mathbf{b_2})_k} = 1$$

We have now found $\frac{\partial z_k}{\partial (\mathbf{W_2})_{kh}}$ and $\frac{\partial z_k}{\partial (\mathbf{b_2})_k}$. If we substitute these into (α) in place of $\frac{\partial z_k}{\partial w}$, which is shown below for reference, we obtain the desired gradients.

$$-\sum_{k=1}^{K} y_k \frac{\partial z_k}{\partial w} + \frac{1}{\sum_{k=1}^{K} \exp(z_k)} \left(\sum_{k=1}^{K} \exp(z_k) \frac{\partial z_k}{\partial w} \right)$$

(c)

Now taking the derivative of z_k with respect to $(\mathbf{W_1})_{hd}$, we obtain:

$$\frac{\partial z_k}{\partial (\mathbf{W_1})_{hd}} = \frac{\partial}{\partial (\mathbf{W_1})_{hd}} \left[\sum_{j=1}^{h} (\mathbf{W_2})_{kj} \; \phi \left(\sum_{m=1}^{d} (\mathbf{W_1})_{jm} \mathbf{x}_m + (\mathbf{b_1})_j \right) + (\mathbf{b_2})_k \right]$$

Note that every term here is 0 but when $j = h$ and $m = d$. Thus:

$$\frac{\partial z_k}{\partial (\mathbf{W_1})_{hd}} = \frac{\partial}{\partial (\mathbf{W_1})_{hd}} \left[(\mathbf{W_2})_{kh} \; \phi \left((\mathbf{W_1})_{hd} \mathbf{x}_d + (\mathbf{b_1})_h \right) + (\mathbf{b_2})_k \right]$$

$$\frac{\partial z_k}{\partial (\mathbf{W_1})_{hd}} = \left[(\mathbf{W_2})_{kh} \; \mathbf{D}\phi \left((\mathbf{W_1})_{hd} \mathbf{x}_d + (\mathbf{b_1})_h \right) \frac{\partial}{\partial (\mathbf{W_1})_{hd}} \left((\mathbf{W_1})_{hd} \mathbf{x}_d + (\mathbf{b_1})_h \right) + \frac{\partial}{\partial (\mathbf{W_1})_{hd}} (\mathbf{b_2})_k \right] \tag{1}$$

Here $\mathbf{D}$ is the derivative operator

$$\frac{\partial z_k}{\partial (\mathbf{W_1})_{hd}} = \left[(\mathbf{W_2})_{kh} \; \mathbf{D}\phi \left((\mathbf{W_1})_{hd} \mathbf{x}_d + (\mathbf{b_1})_h \right) (\mathbf{x}_d) \right]$$

To now differentiate with respect to $(\mathbf{b_1})_h$, just replace $\frac{\partial}{\partial (\mathbf{W_1})_{hd}}$ with $\frac{\partial}{\partial (\mathbf{b_1})_h}$ in label (1)

$$\frac{\partial z_k}{\partial (\mathbf{b_1})_h} = \left[(\mathbf{W_2})_{kh} \; \mathbf{D}\phi \left((\mathbf{W_1})_{hd} \mathbf{x}_d + (\mathbf{b_1})_h \right) \frac{\partial}{\partial (\mathbf{b_1})_h} \left((\mathbf{W_1})_{hd} \mathbf{x}_d + (\mathbf{b_1})_h \right) + \frac{\partial}{\partial (\mathbf{b_1})_h} (\mathbf{b_2})_k \right]$$

$$\frac{\partial z_k}{\partial (\mathbf{b_1})_h} = \left[(\mathbf{W_2})_{kh} \; \mathbf{D}\phi \left((\mathbf{W_1})_{hd} \mathbf{x}_d + (\mathbf{b_1})_h \right) \right]$$

Note that $\phi(a) = \max\{0, a\}$ has a piecewise derivative: 0 when $a \leq 0$ and 1 otherwise. Thus:

$$\frac{\partial z_k}{\partial (\mathbf{W_1})_{hd}} = \begin{cases} [(\mathbf{W_2})_{kh} \, (\mathbf{x}_d)] & \text{When } \left((\mathbf{W_1})_{hd} \mathbf{x}_d + (\mathbf{b_1})_h \right) > 0 \\ 0 & \text{otherwise} \end{cases}$$

Similarly,

$$\frac{\partial z_k}{\partial (\mathbf{b_1})_h} = \begin{cases} [(\mathbf{W_2})_{kh}] & \text{When } \left((\mathbf{W_1})_{hd} \mathbf{x}_d + (\mathbf{b_1})_h \right) > 0 \\ 0 & \text{otherwise} \end{cases}$$

We have now found $\frac{\partial z_k}{\partial (\mathbf{W_1})_{hd}}$ and $\frac{\partial z_k}{\partial (\mathbf{b_1})_h}$. If we substitute these into (α) in place of $\frac{\partial z_k}{\partial w}$, which is shown below for reference, we obtain the desired gradients.

$$-\sum_{k=1}^{K} y_k \frac{\partial z_k}{\partial w} + \frac{1}{\sum_{k=1}^{K} \exp\left(z_k\right)} \left(\sum_{k=1}^{K} \exp\left(z_k\right) \frac{\partial z_k}{\partial w} \right)$$