CS 181 Midterm 1 Solutions
Spring 2015

Name:

1	/ 20
2	/ 10
3	/ 20
4	/ 20
5	/ 10
6	/ 20
Total	/ 100

1. **Logistic Regression [20pts]**

We have data $\{\boldsymbol{x}_n, t_n\}_{n=1}^N$ where $\boldsymbol{x}_n \in \mathbb{R}^D$ and $t_n \in \{0,1\}$, and decide to use logistic regression for this binary classification problem. We use a logistic function $\sigma(z) = 1/(1+\exp\{-z\})$ to determine the conditional probability of a 1 via $\sigma(\boldsymbol{w}^\mathsf{T}\boldsymbol{x})$.

(a) Write the likelihood function for $\boldsymbol{w}$ with these N data.

$$\begin{aligned} L(\boldsymbol{w}; \{\boldsymbol{x}, t\}) &= Pr(\{t\}|\{\boldsymbol{x}\}, \boldsymbol{w}) \\ &= \prod_{n=1}^N Pr(t_n|\boldsymbol{x}_n, \boldsymbol{w}) \\ &= \prod_{n=1}^N \sigma(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_n)^{t_n}(1-\sigma(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_n))^{1-t_n} \end{aligned}$$

(b) Turn this into an "error function" $E(\boldsymbol{w})$ that we might seek to minimize. We would seek to minimize the *negative log likelihood*

$$\begin{aligned} E(\boldsymbol{w}) &= -\log L(\boldsymbol{w}; \{\boldsymbol{x}, t\}) \\ &= -\sum_{n=1}^N t_n \log(\sigma(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_n)) + (1-t_n)\log(1-\sigma(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_n)) \end{aligned}$$

(c) Compute the gradient of $E(\boldsymbol{w})$ in terms of $\boldsymbol{w}$.

Note that $\frac{d}{dz}\sigma(z) = \sigma(z)(1-\sigma(z))$

$$\begin{aligned} \frac{\partial E(\boldsymbol{w})}{\partial \boldsymbol{w}} &= -\sum_{n=1}^N t_n \frac{\partial}{\partial \boldsymbol{w}} \log(\sigma(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_n)) + (1-t_n)\frac{\partial}{\partial \boldsymbol{w}}\log(1-\sigma(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_n)) \\ &= -\sum_{n=1}^N t_n \underbrace{\frac{1}{\sigma(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_n)}}_{\frac{\partial \log\sigma(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_n)}{\partial \sigma(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_n)}} \underbrace{\sigma(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_n)(1-\sigma(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_n))}_{\frac{\partial \sigma(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_n)}{\partial \boldsymbol{w}^\mathsf{T}\boldsymbol{x}_n}} \underbrace{\boldsymbol{x}_n}_{\frac{\partial \boldsymbol{w}^\mathsf{T}\boldsymbol{x}_n}{\partial \boldsymbol{w}}} + \\ &\quad (1-t_n)\underbrace{\frac{1}{1-\sigma(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_n)}}_{\frac{\partial \log(1-\sigma(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_n))}{\partial \sigma(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_n)}} \underbrace{(-\sigma(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_n))(1-\sigma(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_n))}_{\frac{\partial(1-\sigma(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_n))}{\partial \boldsymbol{w}^\mathsf{T}\boldsymbol{x}_n}} \underbrace{\boldsymbol{x}_n}_{\frac{\partial \boldsymbol{w}^\mathsf{T}\boldsymbol{x}_n}{\partial \boldsymbol{w}}} \\ &= -\sum_{n=1}^N t_n(1-\sigma(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_n))\boldsymbol{x}_n - (1-t_n)\sigma(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_n)\boldsymbol{x}_n \\ &= -\sum_{n=1}^N (t_n - \sigma(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}))\boldsymbol{x}_n \end{aligned}$$

(d) To generalize to more than two classes, we can use the softmax function instead of the logistic:

$$\Pr(t_k = 1 \mid \boldsymbol{x}, \{\boldsymbol{w}_{k'}\}_{k'=1}^K) = \frac{\exp\{\boldsymbol{w}_k^\mathsf{T}\boldsymbol{x}\}}{\sum_{k'=1}^K \exp\{\boldsymbol{w}_{k'}^\mathsf{T}\boldsymbol{x}\}}\,.$$

Show that this is not a unique parameterization. That is, show that multiple values of $\{\boldsymbol{w}_k\}_{k=1}^K$ can lead to the same conditional distribution. Suggest constraints on the weights that might resolve this.

To show this is not a unique parameterization, we can construct two sets of weights, denoted $\{\boldsymbol{w}_k\}_{k=1}^K$ and $\{\boldsymbol{v}_k\}_{k=1}^K$, that yield the same probability distribution. Define $\boldsymbol{v}_k = \boldsymbol{w}_k - \boldsymbol{w}_0$, then we have

$$\begin{aligned}
Pr(t_k = 1|\boldsymbol{x}, \{\boldsymbol{v}_{k'}\}_{k'=1}^K) &= \frac{\exp\{\boldsymbol{v}_k^\mathsf{T}\boldsymbol{x}\}}{\sum_{k'=1}^K \exp\{\boldsymbol{v}_{k'}^\mathsf{T}\boldsymbol{x}\}} \\
&= \frac{\exp\{(\boldsymbol{w}_k - \boldsymbol{w}_0)^\mathsf{T}\boldsymbol{x}\}}{\sum_{k'=1}^K \exp\{(\boldsymbol{w}_{k'} - \boldsymbol{w}_0)^\mathsf{T}\boldsymbol{x}\}} \\
&= \frac{\exp\{-\boldsymbol{w}_0^\mathsf{T}\boldsymbol{x}\}\exp\{\boldsymbol{w}_k^\mathsf{T}\boldsymbol{x}\}}{\exp\{-\boldsymbol{w}_0^\mathsf{T}\boldsymbol{x}\}\left(\sum_{k'=1}^K \exp\{\boldsymbol{w}_{k'}^\mathsf{T}\boldsymbol{x}\}\right)} \\
&= Pr(t_k = 1|\boldsymbol{x}, \{\boldsymbol{w}_{k'}\}_{k'=1}^K)
\end{aligned}$$

This reparameterization was suggestive of a constraint: if we set $\boldsymbol{w}_0$ to the zero vector, we get a unique parameterization.

2. **Thresholded Discriminant Functions [10pts]**

Suppose we have the discriminant function $y(\boldsymbol{x}) = \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x} + w_0$, but that rather than assigning $\boldsymbol{x}$ to $\mathcal{C}_1$ when $y(\boldsymbol{x}) \geq 0$ and to $\mathcal{C}_2$ otherwise, we instead assign $\boldsymbol{x}$ to $\mathcal{C}_1$ when $y(\boldsymbol{x}) \geq \eta$ for some η and to $\mathcal{C}_2$ otherwise. Do we gain any generality by moving to this thresholded decision rule? Why or why not?

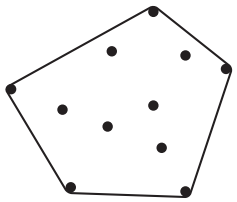
We do not gain any generality because both models describe the same set of classifiers. In particular, a model with $y(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + w_0$ with threshold $y(\boldsymbol{x}) \geq \eta$ is equivalent to the discriminant function $y'(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + (w_0 - \eta)$ with threshold $y'(\mathbf{x}) \geq 0$.

3. **Convex Hulls and Linear Separability [20pts]**

Define the convex hull of a set of data points $\{\boldsymbol{x}_n\}_{n=1}^N$ as the set

$$\left\{ \sum_{n=1}^N \alpha_n \boldsymbol{x}_n \;\middle|\; \alpha_n \geq 0, \sum_{n=1}^N \alpha_n = 1 \right\}.$$

That is, the set of points that are convex sums of $\{\boldsymbol{x}_n\}_{n=1}^N$. In two dimensions a convex hull might look like this:



Show that if two sets of points $\{\boldsymbol{u}_n\}_{n=1}^N$ and $\{\boldsymbol{v}_m\}_{m=1}^M$ are linearly separable, their convex hulls do not intersect.

If $\{\boldsymbol{u}_n\}_{n=1}^N$ and $\{\boldsymbol{v}_m\}_{m=1}^M$ are linearly separable, then there exists a vector $\boldsymbol{w}$ and a w_0 such that

$$\begin{aligned} \boldsymbol{w}^\mathsf{T}\boldsymbol{u}_n + w_0 &\geq 0 \\ \boldsymbol{w}^\mathsf{T}\boldsymbol{v}_m + w_0 &< 0 \end{aligned}$$

for all $\boldsymbol{u}_n$ and $\boldsymbol{v}_m$.

If the convex hulls of $\{\boldsymbol{u}_n\}_{n=1}^N$ and $\{\boldsymbol{v}_m\}_{m=1}^M$ intersect, then there exists a point z on the intersection such that

$$z = \sum_{n=1}^N \alpha_n \boldsymbol{u}_n = \sum_{m=1}^M \beta_m \boldsymbol{v}_m$$

Given $\alpha_n \geq 0, \sum_{n=1}^N \alpha_n = 1$, we have

$$\begin{aligned} \boldsymbol{w}^\mathsf{T} z + w_0 &= \boldsymbol{w}^\mathsf{T} \sum_{n=1}^N \alpha_n \boldsymbol{u}_n + w_0 \\ &= \boldsymbol{w}^\mathsf{T} \sum_{n=1}^N \alpha_n \boldsymbol{u}_n + w_0 \sum_{n=1}^N \alpha_n \\ &= \sum_{n=1}^N \alpha_n (\boldsymbol{w}^\mathsf{T}\boldsymbol{u}_n + w_0) \\ &\geq 0 \end{aligned}$$

Similarly, given $\beta_m \geq 0, \sum_{m=1}^{M} \beta_m = 1$, we have

$$\begin{aligned}
\boldsymbol{w}^{\mathsf{T}} z + w_0 &= \boldsymbol{w}^{\mathsf{T}} \sum_{m=1}^{M} \beta_m \boldsymbol{v}_m + w_0 \\
&= \boldsymbol{w}^{\mathsf{T}} \sum_{m=1}^{M} \beta_m \boldsymbol{v}_m + w_0 \sum_{m=1}^{M} \beta_m \\
&= \sum_{m=1}^{M} \beta_m (\boldsymbol{w}^{\mathsf{T}} \boldsymbol{v}_m + w_0) \\
&< 0
\end{aligned}$$

By contradiction, we proved that for two sets of points that are linearly separable, their convex hulls do not intersect.

4. **Noisy Input Variables in Linear Regression [20pts]**

Consider a linear model of the form

$$y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + \sum_{d=1}^{D} w_d x_d$$

together with a sum-of-squares error function of the form

$$E_D(\boldsymbol{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(\boldsymbol{x}_n, \boldsymbol{w}) - t_n\}^2.$$

Now suppose that Gaussian noise ϵ_n with zero mean and variance σ^2 is added independently to each of the input variables $\boldsymbol{x}_n$. That is, independent noise is added to each dimension of the input. By making use of

$$\mathbb{E}[\epsilon_n] = 0 \qquad\qquad \mathbb{E}[\epsilon_n \epsilon_{n'}] = \begin{cases} \sigma^2 & \text{if } n = n' \\ 0 & \text{if } n \neq n' \end{cases},$$

show that minimizing $E_D(\boldsymbol{w})$ averaged over the noise distribution is equivalent to minimizing the sum-of-squares error for noise-free input variables with the addition of a weight-decay (L_2 norm) regularization term, in which the bias parameter w_0 is omitted from the regularizer.

The question mentions that independent noise is added to each dimension of each input variable $\boldsymbol{x}_n$. That is, our new model becomes

$$\begin{aligned} y'(\boldsymbol{x}_n, \boldsymbol{w}) &= w_0 + \sum_{d=1}^{D} w_d (x_{nd} + \epsilon_{nd}) \\ &= w_0 + \sum_{d=1}^{D} w_d x_{nd} + \sum_{d=1}^{D} w_d \epsilon_{nd} \\ &= y(\boldsymbol{x}_n, \boldsymbol{w}) + \sum_{d=1}^{D} w_d \epsilon_{nd} \end{aligned}$$

where the noise ϵ_{nd} is independent across both the n and d indices. So our new error function is

$$\begin{aligned} E'_D(\boldsymbol{w}) &= \frac{1}{2} \sum_{n=1}^{N} \{y'(\boldsymbol{x}_n, \boldsymbol{w}) - t_n\}^2 \\ &= \frac{1}{2} \sum_{n=1}^{N} \left\{ y(\boldsymbol{x}_n, \boldsymbol{w}) + \sum_{d=1}^{D} w_d \epsilon_{nd} - t_n \right\}^2 \end{aligned}$$

$$= \frac{1}{2} \sum_{n=1}^{N} \left\{ (y(\boldsymbol{x}_n, \boldsymbol{w}) - t_n)^2 + 2(y(\boldsymbol{x}_n, \boldsymbol{w}) - t_n) \left(\sum_{d=1}^{D} w_d \epsilon_{nd} \right) + \left(\sum_{d=1}^{D} w_d \epsilon_{nd} \right)^2 \right\}$$

Taking the expectation of this and using the linearity of expectation, we get

$$\mathbb{E}[E'_D(\boldsymbol{w})] = \frac{1}{2} \sum_{n=1}^{N} \left\{ (y(\boldsymbol{x}_n, \boldsymbol{w}) - t_n)^2 + 2(y(\boldsymbol{x}_n, \boldsymbol{w}) - t_n) \left(\sum_{d=1}^{D} w_d \mathbb{E}[\epsilon_{nd}] \right) + \mathbb{E}\left[\left(\sum_{d=1}^{D} w_d \epsilon_{nd} \right)^2 \right] \right\}.$$

$\mathbb{E}[\epsilon_{nd}]$ is 0, so the second term disappears. Now we look at the third term

$$\begin{aligned}
\mathbb{E}\left[\left(\sum_{d=1}^{D} w_d \epsilon_{nd} \right)^2 \right] &= \mathbb{E}\left[\sum_{d=1}^{D} \sum_{d'=1}^{D} w_d w_{d'} \epsilon_{nd} \epsilon_{nd'} \right] \\
&= \sum_{d=1}^{D} \sum_{d'=1}^{D} w_d w_{d'} \mathbb{E}[\epsilon_{nd} \epsilon_{nd'}] \\
&= \sum_{d=1}^{D} \sum_{d'=1}^{D} w_d w_{d'} \delta_{dd'} \\
&= \sum_{d=1}^{D} w_d^2.
\end{aligned}$$

Using these results, we get

$$\begin{aligned}
\mathbb{E}[E'_D(\boldsymbol{w})] &= \frac{1}{2} \sum_{n=1}^{N} \left\{ (y(\boldsymbol{x}_n, \boldsymbol{w}) - t_n)^2 + \sum_{d=1}^{D} w_d^2 \right\} \\
&= E_D(\boldsymbol{w}) + \frac{N}{2} \sum_{d=1}^{D} w_d^2
\end{aligned}$$

and we see that we get a L_2 regularization term without the bias parameter w_0, as desired.

5. **Multiclass Classification Error Function [10pts]**

Consider a K-class supervised classification scenario with training data $\{\boldsymbol{x}_n, \boldsymbol{t}_n\}_{n=1}^N$, where the $\boldsymbol{t}_n$ are 1-hot binary vectors, with $t_{nk} = 1$ if $\boldsymbol{x}_n$'s true class is k and zero otherwise. Assume we model this problem using a neural network with K outputs that are turned into a conditional probability via a softmax function. That is, the interpretation of the kth output unit, denoted $y_k(\boldsymbol{x}_n, \boldsymbol{w})$ as the probability $p(t_{nk} = 1 \mid \boldsymbol{x}_n, \boldsymbol{w})$. Show that maximizing the conditional likelihood of such a model is equivalent to minimizing the cross-entropy loss function given by

$$E(\boldsymbol{w}) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln y_k(\boldsymbol{x}, \boldsymbol{w}) \,.$$

Let k_n be the class for which $t_{nk} = 1$ for a particular training example n. Then, the likelihood is:

$$\begin{aligned} p(\boldsymbol{t}_n|\boldsymbol{x}_n, \boldsymbol{w}) &= \prod_{n=1}^{N} p(t_{nk_n} = 1|\boldsymbol{x}_n, \boldsymbol{w}) \\ &= \prod_{n=1}^{N} y_{k_n}(\boldsymbol{x}_n, \boldsymbol{w}) \\ &= \prod_{n=1}^{N}\prod_{k=1}^{K} y_k(\boldsymbol{x}_n, \boldsymbol{w})^{t_{nk}}, \end{aligned}$$

where we get the last equality because $t_{nk} = 1$ at exactly k_n and $t_{nk} = 0$ otherwise. Then, the log-likelihood is:

$$\log p(\boldsymbol{t}_n|\boldsymbol{x}_n, \boldsymbol{w}) = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \log y_k(\boldsymbol{x}_n, \boldsymbol{w}) = -E(\boldsymbol{w})$$

So, we have that minimizing cross-entropy loss function is exactly equivalent to maximizing conditional log-likelihood. Since logarithms are monotonically increasing, this is equivalent to maximizing the conditional likelihood.

6. **Tightness of Inequality in Max-Margin Classification [20pts]**

The max-margin classification problem in the linearly separable case is given by

$$\boldsymbol{w}^\star, b^\star = \arg\min_{\boldsymbol{w}} \boldsymbol{w}^\mathsf{T}\boldsymbol{w}$$
$$\text{such that } t_n(\boldsymbol{w}^\mathsf{T}\boldsymbol{\phi}(\boldsymbol{x}_n) + b) \geq 1 \,\forall n\,,$$

where $t_n \in \{-1, +1\}$. Will any of the constraints be tight at the solution? Why or why not? You can explain in words. You may assume that there is at least one example from each class.

Let w^*, b^* minimize $w^t w$ while satisfying the given constraints. Suppose that for w^* and b^*, none of the constraints are tight. Let β be

$$\min_n t_n(w^t\phi(x_n) + b).$$

Note that by assumption $\beta > 1$. Now, consider $w' = \frac{1}{\beta}w^*$, $b' = \frac{1}{\beta}b^*$. Note that

$$w'^t w' = \frac{1}{\beta^2} w^{*t} w^* < w^{*t} w^*$$

and that

$$t_n(w'^t\phi(x_n) + b') = \frac{1}{\beta} t_n(w^{*t}\phi(x_n) + b^*) \geq 1$$

meaning that w', b' give a solution that provides a smaller value for the objective while not violating any constraints. This contradicts that the chosen w^*, b^* minimized the objective while satisfying constraints, and thus at the solution there must exist a tight constraint.

Lastly, we see that

$$\frac{1}{\beta^2} w^{*t} w^* < w^{*t} w^*$$

only holds if $w^* \neq 0$. However, $w^* = 0$ implies that we must have $t_n b \geq 1$ for all n, contradicting that we have at least one example from each class.