Chapter 8

stevenjin8

August 5, 2021

Comments and Proofs

Section 8.3.1

I found section 8.3.1 to be quite confusing, especially equation 8.5:

$$\frac{d}{d\mathbf{w}} f(\mathbf{w}) = \sum_{i=0}^{N} (\mu_i - y_i)\mathbf{x}_i. \tag{1}$$

In equation (1) f is the negative log likelihood and $\mu_i = p(y_i|\mathbf{x}_i, \mathbf{w}) = \sigma(\mathbf{w}^T\mathbf{x})$ where σ is the sigmoid function. We will prove equation (1), but first, some lemmas.

Lemma 1.1. *The derivative of $\sigma(z)$ is $\sigma(z)(1 - \sigma(z))$.*

Proof.

$$\begin{aligned}
\frac{d\sigma}{dz} &= \frac{d}{dz}\frac{1}{1+e^{-z}} \\
&= \frac{1 - 1 + e^{-z}}{(1+e^{-z})^2} \\
&= \frac{1}{1+e^{-z}}\left(1 - \frac{1}{1+e^{-z}}\right) \\
&= \sigma(z)(1-\sigma(z)).
\end{aligned}$$

This result is not trivial, but it is quite intuitive. The whole point of the sigmoid function σ is to monotonicaly map the real line to $(0, 1)$. It follows that when $\sigma(z)$ is extreme, the derivative should be close to 0 , which is exactly what we see. Perhaps, it makes more sense to think of the sigmoid function as the solution to the differential equation:

$$\frac{dy}{dz} = y(y-1).$$

Lemma 1.2. $\sigma(z) + \sigma(-z) = 1$

Proof. Since σ is symmetric about the point $(0, \frac{1}{2})$, we have that $\sigma(z) = 1 - \sigma(-z)$. It follows that $\sigma(z) + \sigma(-z) = 1$.

Now we prove equation (1). Using lemmas 1.1 and 1.2 , we can rewrite the negative log-likelihood as

$$\begin{aligned}\text{NLL}(\mathbf{w}) &= -\sum_{i=1}^{N} y_i \log \mu_i + (1 - y_i) \log(1 - \mu_i) \\ &= -\sum_{i=1}^{N} y_i \log \sigma\left(\mathbf{w}^T \mathbf{x}_i\right) + (1 - y_i) \log \sigma\left(-\mathbf{w}^T \mathbf{x}_i\right)\end{aligned}$$

Now, we find the derivative with respect to $\mathbf{w}$:

$$\begin{aligned}\frac{d\text{NLL}}{d\mathbf{w}} &= -\sum_{i=1}^{N} y_i \frac{d}{d\mathbf{w}} \log \sigma(\mathbf{w}^T \mathbf{x}_i) + (1 - y_i) \frac{d}{d\mathbf{w}} \log \sigma(-\mathbf{w}^T \mathbf{x}_i) \\ &= \sum y_i (1 - \sigma(\mathbf{w}^T \mathbf{x}_i)) \mathbf{x}_i - (1 - y_i) \sigma(\mathbf{w}^T \mathbf{x}_i) \mathbf{x}_i \\ &= \sum (\sigma(\mathbf{w}^T \mathbf{x}_i) - y_i) \mathbf{x}_i \\ &= \sum (\mu_i - y_i) \mathbf{x}_i.\end{aligned}$$

Given lemma 1.1 and 1.2, this result comes naturally, but it is still a good exercise to do since the proof for the backpropagation algorithm is similar.

Section 8.3.3

I had a lot of trouble with equation 8.15 and 8.16. I kept mixing up θ and θ_k. Also, keep in mind that the Hessian matrix $\mathbf{H}_k$ is symmetric due to the law of mixed partials.

Section 8.6.2

I found this first paragraph of this section extremely confusing. Initially, I thought that $\mathbf{x}_i$ was the ith data point and $r_i \in \{0, 1\}$ indicated whether $\mathbf{x}_i$ was observed. I think the author meant that given a data point $\mathbf{x}$, the variable $r_i \in \{0, 1\}$ indicates whether the ith feature of $\mathbf{x}$ was observed. This section would make a lot more sense if each $\mathbf{x}_i$ was replaced with x_i.

Exercises

Exercise 3

a. See section 8.3.1 above.
b. See section 8.3.1 above.

c. Let $\mathbf{H}$ be the Hessian matrix of a continuous twice-differentiable function $f(\mathbf{x})$. Recall that, $H_{i,j} = \frac{\partial^2 f}{\partial x_i \partial x_j}$. By the law of mixed partials, $H_{i,j} = H_{j,i}$. In other words, $\mathbf{H}$ is symmetric. It follows that $\mathbf{H}$ has D eigenvalues $\lambda_1, \lambda_2, ..., \lambda_D$. Since eigenvectors of different eigenvalues are linearly independent, there exists a orthonormal eigenbasis $\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_D$ where $\mathbf{H}\mathbf{p}_i = \lambda_i \mathbf{p}_i$. If follows that

$$\mathbf{H} = \mathbf{P}\mathbf{D}\mathbf{P}^T,$$

where $\mathbf{D} = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_D)$ and $\mathbf{P} = [\mathbf{p}_1 \ \ \mathbf{p}_2 \ \ ... \ \ \mathbf{p}_D]$.

Now we prove that all eigenvalues are positive. Following equation (2), we have

$$\mathbf{D} = \mathbf{P}^T \mathbf{H} \mathbf{P}.$$

The ith eigenvalue is the ith element of

$$\begin{aligned}
\lambda_i \mathbf{e}_i &= \mathbf{D}\mathbf{e}_i \\
&= \mathbf{P}^T \mathbf{H} \mathbf{P} \mathbf{e}_i \\
&= \mathbf{P}^T \mathbf{H} \mathbf{p}_i.
\end{aligned}$$

The ith eigenvalue then is given by

$$\begin{aligned}
\lambda_i &= \mathbf{p}_i^T \mathbf{H} \mathbf{p}_i \\
&= \mathbf{p}_i^T \mathbf{X}^T \mathbf{S} \mathbf{X} \mathbf{p}_i \\
&= (\mathbf{X}\mathbf{p}_i)^T \mathbf{S} (\mathbf{X}\mathbf{p}_i).
\end{aligned}$$

If we let $\mathbf{a}_i = \mathbf{X}\mathbf{p}_i$, we have

$$\begin{aligned}
\lambda_i &= \mathbf{a}_i^T \mathbf{S} \mathbf{a}_i \\
&= \sum_{j=1}^{D} a_{ij}^2 \mu_j (1 - \mu_j) \\
&> 0.
\end{aligned}$$

The strict inequality comes from the fact that $\mathbf{X}$ is full rank and $\mathbf{p}_i$ is non-zero. Thus, at least one element of $\mathbf{a}_i = \mathbf{X}\mathbf{p}_i$ is non-zero. Since $\mathbf{H}$ is a symmetric matrix with positive eigenvalues, it is positive definite.

Exercise 5

In this exercise we show that $\sum_{c=1}^{C} \hat{w}_{cj} = 0$ for any feature j when maximizing

$$f(\mathbf{W}) = \sum_{i=1}^{N} p(y_i | x_i, \mathbf{W}) - \sum_{c=1}^{C} \|\mathbf{w}_c\|_2^2.$$

Lemma 1.3. $p(y|\mathbf{x}, \mathbf{W}) = p(y|\mathbf{x}, \mathbf{W} + \mathbf{A})$ *where* $\mathbf{A}$ *is any matrix in the form* $\mathbf{A} = [\mathbf{0}\ \ a_1\mathbf{1}\ \ \ldots\ \ a_D\mathbf{1}]$.

Let $\mathbf{a} = (a_1, \ldots, a_D)$. It follows that

$$\begin{aligned} p(y = c|\mathbf{x}, \mathbf{W} + \mathbf{A}) &= \frac{\exp(w_{c0} + (\mathbf{w}_c + \mathbf{a})^T\mathbf{x})}{\sum_{c'=1}^{C} \exp(w_{c'0} + (\mathbf{w}_{c'} + \mathbf{a})^T\mathbf{x})} \\ &= \frac{\exp(w_{c0} + \mathbf{w}_c^T\mathbf{x}) \exp(\mathbf{a}^T\mathbf{x})}{\sum_{c'=1}^{C} \exp(w_{c'0} + \mathbf{w}_{c'}^T\mathbf{x}) \exp(\mathbf{a}^T\mathbf{x})} \\ &= \frac{\exp(w_{c0} + \mathbf{w}_c^T\mathbf{x})}{\sum_{c'=1}^{C} \exp(w_{c'0} + \mathbf{w}_{c'}^T\mathbf{x})} \\ &= p(y = c|\mathbf{x}, \mathbf{W}). \end{aligned}$$

Now we finish the exercise with a proof by contradiction. Say $\mathbf{W}$ maximizes f and $\sum_{c=1}^{C} \hat{w}_{cj} \neq 0$ for one or more j. Let $\mathbf{W}' = \mathbf{W} - \mathbf{A}$ where $\mathbf{A} = [\mathbf{0}\ \ a_1\mathbf{1}\ \ \ldots\ \ a_D\mathbf{1}]$ and $a_i = \frac{1}{C}\sum_{c=1}^{C} w_{c,i}$. In other words, $\mathbf{X}'$ is $\mathbf{X}$ with centered columns (except for the first column). By lemma 1.3, we have

$$\sum_{i=0}^{N} p(y_i|\mathbf{x}_i, \mathbf{W}) = \sum_{i=0}^{N} p(y_i|\mathbf{x}_i, \mathbf{W}').$$

Since $\mathbf{X}'$ is more centered, we also have

$$\sum_{c=1}^{C} \|\mathbf{w}_c\|_2^2 > \sum_{c=1}^{C} \|\mathbf{w}'_c\|_2^2,$$

but equations 2 and 3 imply that $f(\mathbf{W}) < f(\mathbf{W}')$, contradicting our initial statement that $\mathbf{W}$ maximizes f. It follows that if $\hat{\mathbf{W}}$ maximizes f, then $\sum_{c=1}^{C} \hat{w}_{cj} = 0$ for any $j > 0$.

Exercise 6

a. *True.* By exercise 3, $\mathbf{H}$, the Hessian of J, is positive definite. It follows that J is convex and has a single local optimum. This is the multivariate equivalent of a function always having a positive second derivative.
b. *False.* Since $\frac{\partial}{\partial x_i}\|\mathbf{w}\|_2^2 = 0$ when $x_i = 0$, we know that $\frac{\partial J}{\partial w_i}|_{w_i=0} = 0$ if and only if $\frac{\partial \ell}{\partial x_i}|_{w_i=0} = 0$, which is very unlikely.
c. *True.* Since $\lambda = 0$, minimizing J is equivalent to maximizing ℓ. Since the data are linearly separable, there exists weights $\mathbf{w} \neq \mathbf{0}$ such that our model always makes the right prediction. In other words, $y_i\mathbf{x}_i^T\mathbf{w} > 0$. Now, consider $\ell(a\mathbf{w}, \mathcal{D})$ for some $a > 1$. Since σ is positive monotonic, $\sigma(y_i\mathbf{x}_i^T(a\mathbf{w})) > \sigma(y_i\mathbf{x}_i^T\mathbf{w})$ and $\ell(a\mathbf{w}, \mathcal{D}) > \ell(\mathbf{w}, \mathcal{D})$. It follows all the non-zero weights will become infinite.
d. *False.* No because as λ grows, $\hat{\mathbf{w}}$ will start underfitting to the data.
e. *False.* No because as λ grows, $\hat{\mathbf{w}}$ will start underfitting to the data.

Exercise 7

a. The decision boundary is around $X_2 = 3(X_1 - 3)$. No errors on the training set as the data are linearly separable.
b. The decision boundary is around $X_1 = X_2$ and the misclassification rate is $1/13$.
c. Since we are heavily regularizing w_1, $\hat{w}_1 \approx 0$ and our prediction will only be affected by X_2. A decision boundary is $X_2 = 3$, and the misclassification rate is $2/13$.
d. Since we are heavily regularizing w_2, $\hat{w}_2 \approx 0$ and our prediction will only be affected by X_1. A decision boundary is $X_2 = 5$, and the misclassification rate is $0/13$.