6.867: Exercises (Week 1)

Sept 15, 2017

1. (Bishop 3.1) Show that the tanh function and the logistic sigmoid function σ are related by

$$\tanh(a) = 2\sigma(2a) - 1 \tag{1}$$

Hence show that a general linear combination of logistic sigmoid functions of the form

$$y(x, w) = w_0 + \sum_{j=1}^{M} w_j \sigma(\frac{x - u_j}{s}) \tag{2}$$

is equivalent to a linear combination of tanh functions of the form

$$y(x, b) = b_0 + \sum_{j=1}^{M} b_j \tanh(\frac{x - u_j}{2s}) \tag{3}$$

and find expressions to relate the new parameters $\{b_0, \dots, b_M\}$ to the original parameters $\{w_0, \dots, w_M\}$.

Solution: Since $\sigma(a) = \frac{1}{1+\exp(-a)}$, we have

$$\begin{aligned} 2\sigma(2a) - 1 &= \frac{2}{1 + e^{-2a}} - 1 \\ &= \frac{2}{1 + e^{-2a}} - \frac{1 + e^{-2a}}{1 + e^{-2a}} \\ &= \frac{1 - e^{-2a}}{1 + e^{-2a}} \\ &= \frac{e^{a} - e^{-a}}{e^{a} + e^{-a}} \\ &= \tanh(a) \end{aligned}$$

Let $a_j = (x - u_j)/2s$. We can rewrite (2) as

$$\begin{aligned} y(x, w) &= w_0 + \sum_{j=1}^{M} w_j \sigma(2a_j) \\ &= w_0 + \sum_{j=1}^{M} \frac{w_j}{2}(2\sigma(2a_j) - 1 + 1) \\ &= b_0 + \sum_{j=1}^{M} b_j \tanh(a_j), \end{aligned}$$

where $b_j = w_j/2$ for $j = 1, \ldots, M$, and $b_0 = w_0 + \sum_{j=1}^{M} w_j/2$.

2. (Bishop 3.2) Show that the matrix
$$\Phi(\Phi^\top \Phi)^{-1}\Phi^\top \tag{4}$$
takes any vector v and projects it onto the space spanned by the columns of Φ. Use this result to show that the least-squares solution ($f = \Phi w^*$, where $w^* = (\Phi^\top\Phi)^{-1}\Phi^\top Y$) corresponds to an *orthogonal* projection of the target vector Y onto the subspace spanned by the columns of Φ.

Solution: We first write
$$\begin{aligned}\Phi(\Phi^\top\Phi)^{-1}\Phi^\top v &= \Phi\tilde{v} \\ &= \phi_1\tilde{v}^{(1)} + \phi_2\tilde{v}^{(2)} + \cdots + \phi_M\tilde{v}^{(M)}\end{aligned} \tag{5}$$
where ϕ_m is the m-th column of Φ, $\tilde{v} = (\Phi^\top\Phi)^{-1}\Phi^\top v$, and $\tilde{v}^{(m)}$ is the m-th element of the vector $\tilde{v}$. There, $\Phi(\Phi^\top\Phi)^{-1}\Phi^\top v$ can be represented as a linear combination of all the columns of Φ, which implies that $\Phi(\Phi^\top\Phi)^{-1}\Phi^\top v$ is a projection of v to the column space of Φ.

By comparing with the least squares solution, we see that $f = \Phi w^* = \Phi(\Phi^\top\Phi)^{-1}\Phi^\top Y$ corresponds to a projection of Y onto the space spanned by the columns of Φ. To see that this is indeed an orthogonal projection, here are two alternative solutions:

(1) We first note that for any column of Φ, ϕ_j, we have
$$\Phi(\Phi^\top\Phi)^{-1}\Phi^\top\phi_j = [\Phi(\Phi^\top\Phi)^{-1}\Phi^\top\Phi]_j = \phi_j \tag{6}$$
and therefore,
$$\begin{aligned}(f - Y)^\top\phi_j = (\Phi w^* - Y)^\top\phi_j &= Y^\top(\Phi(\Phi^\top\Phi)^{-1}\Phi^\top - I)^\top\phi_j \\ &= Y^\top(\Phi(\Phi^\top\Phi)^{-1}\Phi^\top - I)\phi_j \\ &= Y^\top(\Phi(\Phi^\top\Phi)^{-1}\Phi^\top\phi_j - \phi_j) \\ &= 0\end{aligned} \tag{7}$$
Thus, $(f - Y)$ is orthogonal to every column of Φ, which implies that it is orthogonal to the column space of Φ.

(2) Suppose that this is indeed an orthogonal projection, then by definition, $(Y - f)^\top q = 0$ for any vector q in the column space of Φ. Let us prove this by contradiction. Suppose that it is not, then there exist a vector $\tilde{q}$ in the column space of Φ such that $(Y - f)^\top\tilde{q} \neq 0$. Without loss of generality, assume that $(Y - f)^\top\tilde{q} > 0$ and $\|\tilde{q}\|_2 = 1$. Let us consider $\tilde{f} = f + \delta\tilde{q}$ for some $\delta > 0$, and the sum-of-squares error of $\tilde{f}$:
$$\begin{aligned}(Y - \tilde{f})^\top(Y - \tilde{f}) &= (Y - f - \delta\tilde{q})^\top(Y - f - \delta\tilde{q}) \\ &= (Y - f)^\top(Y - f) + \delta^2\tilde{q}^\top\tilde{q} - 2\delta(Y - f)^\top\tilde{q} \\ &= (Y - f)^\top(Y - f) + \delta^2 - 2\delta(Y - f)^\top\tilde{q}\end{aligned}$$

Let us consider the term $\delta^2 - 2\delta(Y - f)^T\tilde{q}$. By assumption, $(Y - f)^T\tilde{q} > 0$. Furthermore, for δ small enough, the linear term $2\delta(Y - f)^T\tilde{q}$ will dominate the quadratic term δ^2. In other words, for $\delta > 0$ small enough, we have that $\delta^2 - 2\delta(Y - f)^T\tilde{q} < 0$, which implies that

$$(Y - \tilde{f})^T(Y - \tilde{f}) < (Y - f)^T(Y - f), \text{for small enough } \delta.$$

However, this implies that $\tilde{f}$ achieves a smaller sum-of-square error than f, which contradicts the fact that f is the least squares solution. Therefore, f must correspond to a projection of Y onto the column space of Φ.

3. (Bishop 3.3) Consider a dataset in which each data point (x_n, y_n) is associated with a weighting factor $r_n > 0$, so that the sum-of-squares error function becomes

$$E_D(w) = \frac{1}{2}\sum_{n=1}^{N} r_n\{y_n - w^T\phi(x_n)\}^2 \tag{8}$$

Find an expression for the solution w^* that minimizes the sum-of-squares error. Give two alternative interpretations of the weighted sum-of-squares error function in terms of (i) data dependent noise variance and (ii) replicated data points.

Solution: If we define $R = \text{diag}(r_1, \ldots, r_N)$ to be a diagonal matrix containing the weighting coefficients, then we can write the weighted sum-of-squares cost function in the form

$$E_D(w) = \frac{1}{2}(Y - \Phi w)^T R(Y - \Phi W).$$

Setting the derivative with respect to w to zero, and then we obtain

$$w^* = (\Phi^T R\Phi)^{-1}\Phi^T RY$$

which reduces to the standard solution for the case $R = I$.

If we compare the sum-of-squares error function to the log likelihood function (see Lecture 2 slides), we see that r_n can be regarded as the inverse variance, particular to the data point (x_n, y_n). Alternatively, r_n can be regarded as an *effective* number of replicated observations of data point (x_n, y_n); this becomes particularly clear if r_n taking positive integer values, although it is valid for any $r_n > 0$.

4. (Bishop 3.4) Consider a linear model of the form

$$f(x, w) = w_0 + \sum_{i=1}^{D} w_i x^{(i)} \tag{9}$$

where $x^{(i)}$ is the i-th coordinate of the vector x, and together with a sum-of-squares error function of the form

$$E_D(w) = \frac{1}{2}\sum_{n=1}^{N}\{y_n - f(x_n, w)\}^2 \tag{10}$$

Now suppose that Gaussian noise ϵ_i with zero mean and variance σ^2 is added independently to each of the input variables $x^{(i)}$. By making use of $\mathbb{E}[\epsilon_i] = 0$ and $\mathbb{E}[\epsilon_i\epsilon_j] = \delta_{ij}\sigma^2$, show that minimizing E_D 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 regularization term, in which the bias parameters w_0 is omitted from the regularizer.

Solution: Let

$$\tilde{y}_n = w_0 + \sum_{i=1}^{D} w_i(x_n^i + \epsilon_{ni}) = f_n + \sum_{i=1}^{D} w_i\epsilon_{ni}$$

where $f_n = f(x_n, w)$ is the predicted value for the n-th data point and $\epsilon_{ni} \sim \mathcal{N}(0, \sigma^2)$. From (10), we then define

$$\begin{aligned}
\tilde{E} &= \frac{1}{2}\sum_{n=1}^{N}\{y_n - \tilde{y}_n\}^2 \\
&= \frac{1}{2}\sum_{n=1}^{N}\{\tilde{y}_n^2 - 2\tilde{y}_n y_n + y_n^2\} \\
&= \frac{1}{2}\sum_{n=1}^{N}\left\{f_n^2 + 2f_n\sum_{i=1}^{D} w_i\epsilon_{ni} + \Big(\sum_{i=1}^{D} w_i\epsilon_{ni}\Big)^2 - 2y_n f_n - 2y_n\sum_{i=1}^{D} w_i\epsilon_{ni} + y_n^2\right\}
\end{aligned} \tag{11}$$

If we take the expectation of $\tilde{E}$ under the distribution of ϵ_{ni}, we see that the second and fifth terms disappear, since $\mathcal{E}[\epsilon_{ni}] = 0$. For the third term we get

$$\mathbb{E}\left[\Big(\sum_{i=1}^{D} w_i\epsilon_{ni}\Big)^2\right] = \sum_{i=1}^{D} w_i^2\sigma^2$$

since the ϵ_{ni} are all independent with variance σ^2. From this and (10), we see that

$$\mathbb{E}[\tilde{E}] = E_D + \frac{N}{2}\sum_{i=1}^{D} w_i^2\sigma^2$$

as required.

5. (Bishop 3.5) Using the technique of Lagrange multipliers (Appendix E of Bishop if you are not familiar with), show that minimization of the regularized error function

$$\frac{1}{2}\sum_{n=1}^{N}\{y_n - w^T\phi(x_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q \tag{12}$$

is equivalent to minimizing the unregularized sum-of-squares error

$$E_D(w) = \frac{1}{2}\sum_{n=1}^{N}\{y_n - w^T\phi(x_n)\}^2 \tag{13}$$

subject to the constraint

$$\sum_{j=1}^{M}|w_j|^q \leqslant \eta \tag{14}$$

Discuss the relationship between the parameters η and λ.

Solution: We can rewrite the constraint (14) as

$$\frac{1}{2}(\sum_{j=1}^{M}|w_j|^q - \eta) \leqslant 0$$

where we have incorporated the 1/2 scaling factor for convenience. Clearly this does not affect the constraint.

Employing the technique of Lagrange multipliers, we can combine the condition with (13) to obtain the Lagrangian function

$$L(w,\lambda) = \frac{1}{2}\sum_{n=1}^{N}\{y_n - w^T\phi(x_n)\}^2 + \frac{\lambda}{2}(\sum_{j=1}^{M}|w_j|^q - \eta) \tag{15}$$

and by comparing this with (12), we see immediately that they are identical in their dependence on w.

Now suppose we choose a specific value of $\lambda > 0$ and minimize (12). Denoting the resulting value of w by $w^*(\lambda)$, and using the KKT condition, we see that the value of η is given by

$$\eta = \sum_{j=1}^{M}|w_j^*(\lambda)|^q.$$

6. (Bishop 3.6, Modified) Consider a linear basis function regression model for a multivariate target variable y (i.e. y is a column vector) having a Gaussian distribution of the form

$$p(y|W,\Sigma) = \mathcal{N}(f(x,W),\Sigma) \tag{16}$$

where $f(x,W) = W^T\phi(x)$, together with a training dataset comprising input basis vectors $\phi(x_n)$ and corresponding target vectors y_n, with $n = 1,\ldots,N$.

1. Write down the log likelihood function given the data.
2. Derive the maximum likelihood estimator W_{ML} for the parameter matrix W.

3. The maximum likelihood estimator for the covariance matrix Σ_{ML} involves optimization over positive definite matrices, and is very complex. However, as you see in Lectures, the maximum likelihood estimator often takes an intuitive form. Based on W_{ML} from (2) and your experience when y_n is a scalar, guess Σ_{ML}.

Solution: (1) We first write down the log likelihood function which is given by

$$\ln L(W, \Sigma) = -\frac{N}{2} \ln |\Sigma| - \frac{1}{2} \sum_{n=1}^{N} (y_n - W^T \phi(x_n))^T \Sigma^{-1} (y_n - W^T \phi(x_n))$$

(2) We set the derivative with respect to W equal to zero, giving

$$0 = \sum_{n=1}^{N} \Sigma^{-1} (y_n - W^T \phi(x_n)) \phi(x_n)^T$$

Multiplying through by Σ and introducing the design matrix Φ and the target data matrix T (i.e., the ith row of T is the vector y_i^T), we have

$$\Phi^T \Phi W = \Phi^T T$$

Solving for W then gives $W_{ML} = (\Phi^T \Phi)^{-1} \Phi^T T$.

(3) Σ_{ML} takes the following intuitive form:

$$\Sigma_{ML} = \frac{1}{N} \sum_{n=1}^{N} (y_n - W_{ML}^T \phi(x_n))(y_n - W_{ML}^T \phi(x_n))^T.$$

7. (JWHT 3.5, Modified) Consider a dataset with N data points, $(x_1, y_1), \ldots, (x_N, y_N)$, where both x_n and y_n are scalar numbers. Consider the fitted values that result from performing linear regression without an intercept. In this setting, the ith fitted value takes the form

$$f(x_i, w) = x_i w$$

where $w \in \mathbb{R}$. Derive the w^* that minimizes the sum-of-squares error. Show that we can write

$$f(x_i, w) = \sum_{j=1}^{N} a_j y_j$$

and derive the equation for a_j.

(Note: We interpret this result by saying that the fitted values from linear regression are linear combinations of the target values.)

Solution: The sum-of-squares error is

$$\frac{1}{2}\sum_{i=1}^{N}(y_i - x_i w)^2$$

Setting the derivative with respect to w equal to zero, and we obtain

$$\sum_{i=1}^{N}(y_i - x_i w)x_i = 0 \Rightarrow w = (\sum_{i=1}^{N} x_i y_i)/(\sum_{k=1}^{N} x_k^2)$$

Then,

$$f(x_i, w) = \left[(\sum_{j=1}^{N} x_j y_j)/(\sum_{k=1}^{N} x_k^2)\right] x_i = \sum_{j=1}^{N} \frac{x_j x_i}{\sum_{k=1}^{N} x_k^2} y_j$$

which implies that $a_j = \frac{x_j x_i}{\sum_{k=1}^{N} x_k^2}$.

8. We have provided the advertisement data used in lectures. To gain hands on experience, you are highly encouraged to build your own regression model with the data. As a starting point, you could build the same model as in lectures and check your understanding with the results in the lecture slides.