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 $\{b_0, \ldots, b_M\}$ to the original parameters $\{w_0, \ldots, w_M\}$.

2. (Bishop 3.2) Show that the matrix

$$\Phi(\Phi^T\Phi)^{-1}\Phi^T \tag{4}$$

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 $w^* = (\Phi^T\Phi)^{-1}\Phi^T Y$) corresponds to an *orthogonal* projection of the target vector Y onto the subspace spanned by the columns 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.

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.

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.

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}.

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.)

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.