

iris plant setosa versicolor virginica

Use the dataset `Iris-cleaned.csv` to predict plant species from sepal length and sepal width measurements.

One of the problems of using linear regression to solve classification problems like the Iris problem described above is that the predictions, $\hat{\mathbf{Y}}$, of linear regression are not true probability distributions like one-hot-encoding. We can fix this problem by using a special function called the **softmax function** that converts the output, $\hat{\mathbf{Y}}$, of linear regression into a genuine probability distribution, $\hat{\mathbf{P}}$. The resulting classifier is called **logistic regression**. Unfortunately, logistic regression, is nonlinear, so the normal equations can no long be applied.

40 Definition (Linearity)
A function $\mathbf{y} = f(\mathbf{x})$ with n inputs and m outputs is is linear if there exits an $m \times n$ matrix $\mathbf{A}$ such that $f(\mathbf{x}) = \mathbf{Ax}$.

Note that a characteristic of linear function is that $f(\mathbf{0}) = \mathbf{0}$.

---

41 Definition (Training vs Inference)
Linear regression networks have two functional forms:

$$\begin{array}{rl} \text{Training} & \hat{\mathbf{y}} = f_{\mathbf{X}}(\mathbf{w}) = \mathbf{Xw} \\ \text{Inference} & \hat{\mathbf{y}} = f_{\mathbf{w}}(\mathbf{X}) = \mathbf{Xw} \end{array}$$

During training, the data matrix $\mathbf{X}$, is constant while during inference, the weights $\mathbf{w}$ are constant.

---

42 Definition (Linear Combination)
Linear regression can be interpreted as a **linear combination** of features plus a bias.

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{Xw} \\ &= \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_d & \mathbf{1} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \\ b \end{pmatrix} \\ &= w_1\mathbf{x}_1 + w_2\mathbf{x}_2 + \cdots + w_d\mathbf{x}_d + b\mathbf{1} \end{aligned}$$

The predictions for the target, $\hat{\mathbf{y}}$, is just a weighted sums of the d features (columns of $\mathbf{X}$) plus a bias.

For the case of r targets we have that $\hat{\mathbf{Y}}$ and $\hat{\mathbf{W}}$ are matrices where

$$\begin{aligned} \hat{\mathbf{Y}} &= \mathbf{XW} \\ \begin{pmatrix} \mathbf{y}_1 & \mathbf{y}_2 & \cdots & \mathbf{y}_r \end{pmatrix} &= \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_d & \mathbf{1} \end{pmatrix} \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1r} \\ w_{21} & w_{22} & \cdots & w_{2r} \\ \vdots & \vdots & & \vdots \\ w_{d1} & w_{d2} & \cdots & w_{dr} \\ b_1 & b_2 & \cdots & b_r \end{pmatrix} \end{aligned}$$

---

3 Optimization

43 Definition (Increasing/Decreasing Function)
A function is increasing if its derivative is positive and it is decreasing if its derivative is negative.

---

44 Example (Minimum of a Function)

(a) Explain why the minimum value of a smooth function can only occur when the derivative of the function equals zero.

(b) If the derivative of a smooth function equals zero, must the function value be at either a maximum or minimum value? Explain.
