رگرسیون لجستیک

Logistic Regression

مجموعه ای از d پیش‌بینی کننده یا متغیر مستقل $x_1, x_2, \ldots, x_d$ و یک متغیر پاسخ دوتایی یا برنولی Y که فقط دو مقدار می‌گیرد.

$$\tilde{\mathbf{x}}_i = (1, x_1, x_2, \cdots, x_d)^T \in \mathbb{R}^{d+1}$$

$$P(Y=1|\tilde{\mathbf{X}}=\tilde{\mathbf{x}}) = \pi(\tilde{\mathbf{x}}) \qquad P(Y=0|\tilde{\mathbf{X}}=\tilde{\mathbf{x}}) = 1-\pi(\tilde{\mathbf{x}})$$

که در آن $\pi(\tilde{\boldsymbol{x}})$ مقدار واقعی پارامتر و برای ما ناشناخته است.

تابع منطقی یا هلالی (Logistic or Sigmoid)

$$\boxed{\theta(z) = \frac{1}{1+\exp\{-z\}} = \frac{\exp\{z\}}{1+\exp\{z\}}}$$

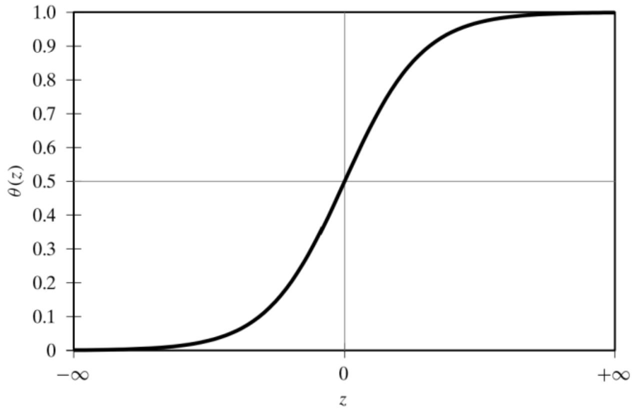


Figure 24.1. Logistic function.

$$f(\tilde{\mathbf{x}}) = \omega_0 \cdot x_0 + \omega_1 \cdot x_1 + \omega_2 \cdot x_2 + \cdots + \omega_d \cdot x_d = \tilde{\boldsymbol{\omega}}^T \tilde{\mathbf{x}}$$

$$P(Y=1|\,\tilde{\mathbf{X}}=\tilde{\mathbf{x}}) = \pi(\tilde{\mathbf{x}}) = \theta(f(\tilde{\mathbf{x}})) = \theta(\tilde{\boldsymbol{\omega}}^T \tilde{\mathbf{x}}) = \frac{\exp\{\tilde{\boldsymbol{\omega}}^T \tilde{\mathbf{x}}\}}{1+\exp\{\tilde{\boldsymbol{\omega}}^T \tilde{\mathbf{x}}\}}$$

$$P(Y=0|\,\tilde{\mathbf{X}}=\tilde{\mathbf{x}}) = 1 - P(Y=1|\,\tilde{\mathbf{X}}=\tilde{\mathbf{x}}) = \theta(-\tilde{\boldsymbol{\omega}}^T \tilde{\mathbf{x}}) = \frac{1}{1+\exp\{\tilde{\boldsymbol{\omega}}^T \tilde{\mathbf{x}}\}}$$

$$\boxed{P(Y|\,\tilde{\mathbf{X}}=\tilde{\mathbf{x}}) = \theta(\tilde{\boldsymbol{\omega}}^T \tilde{\mathbf{x}})^Y \cdot \theta(-\tilde{\boldsymbol{\omega}}^T \tilde{\mathbf{x}})^{1-Y}}$$

Log-Odds Ratio

$$\ln\big(\mathrm{odds}(Y=1|\tilde{\mathbf{X}}=\tilde{\mathbf{x}})\big) = \ln\left(\frac{P(Y=1|\tilde{\mathbf{X}}=\tilde{\mathbf{x}})}{1-P(Y=1|\tilde{\mathbf{X}}=\tilde{\mathbf{x}})}\right) = \ln\left(\exp\{\tilde{\boldsymbol{\omega}}^T \tilde{\mathbf{x}}\}\right) = \tilde{\boldsymbol{\omega}}^T \tilde{\mathbf{x}}$$

$$= \omega_0 \cdot x_0 + \omega_1 \cdot x_1 + \cdots + \omega_d \cdot x_d$$

$$\mathrm{logit}(z) = \ln\left(\frac{z}{1-z}\right)$$

$$\ln\big(\mathrm{odds}(Y=1|\tilde{\mathbf{X}}=\tilde{\mathbf{x}})\big) = \mathrm{logit}\big(P(Y=1|\tilde{\mathbf{X}}=\tilde{\mathbf{x}})\big)$$

رگرسیون منطق دوتایی (Binary Logistic Regression)

Algorithm 24.1: Logistic Regression: Stochastic Gradient Ascent

LogisticRegression-SGA $(\mathbf{D}, \eta, \epsilon)$:

foreach $\mathbf{x}_i \in \mathbf{D}$ **do** $\tilde{\mathbf{x}}_i^T \leftarrow \begin{pmatrix} 1 & \mathbf{x}_i^T \end{pmatrix}$ // map to $\mathbb{R}^{d+1}$

$t \leftarrow 0$ // step/iteration counter

$\tilde{\mathbf{w}}^0 \leftarrow (0, \ldots, 0)^T \in \mathbb{R}^{d+1}$ // initial weight vector

repeat

  $\tilde{\mathbf{w}} \leftarrow \tilde{\mathbf{w}}^t$ // make a copy of $\tilde{\mathbf{w}}^t$

  foreach $\tilde{\mathbf{x}}_i \in \tilde{\mathbf{D}}$ *in random order* **do**

    $\nabla(\tilde{\mathbf{w}}, \tilde{\mathbf{x}}_i) \leftarrow \left(y_i - \theta(\tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_i)\right) \cdot \tilde{\mathbf{x}}_i$ // compute gradient at $\tilde{\mathbf{x}}_i$

    $\tilde{\mathbf{w}} \leftarrow \tilde{\mathbf{w}} + \eta \cdot \nabla(\tilde{\mathbf{w}}, \tilde{\mathbf{x}}_i)$ // update estimate for $\tilde{\mathbf{w}}$

  $\tilde{\mathbf{w}}^{t+1} \leftarrow \tilde{\mathbf{w}}$ // update $\tilde{\mathbf{w}}^{t+1}$

  $t \leftarrow t + 1$

until $\|\tilde{\mathbf{w}}^t - \tilde{\mathbf{w}}^{t-1}\| \leq \epsilon$

$$\boxed{\tilde{\mathbf{w}}^{t+1} = \tilde{\mathbf{w}}^t + \eta \cdot \nabla(\tilde{\mathbf{w}}^t)}$$

Stochastic Gradient Ascent

$$\boxed{\nabla(\tilde{\mathbf{w}}, \tilde{\mathbf{x}}_i) = \left(y_i - \theta(\tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_i)\right) \cdot \tilde{\mathbf{x}}_i}$$

برآوردگر بیشینه درستنمایی (Maximum Likelihood Estimation)

$$L(\tilde{\mathbf{w}}) = P(Y|\tilde{\mathbf{w}}) = \prod_{i=1}^{n} P(y_i \mid \tilde{\mathbf{x}}_i) = \prod_{i=1}^{n} \theta(\tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_i)^{y_i} \cdot \theta(-\tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_i)^{1-y_i}$$

$$\boxed{\ln\left(L(\tilde{\mathbf{w}})\right) = \sum_{i=1}^{n} y_i \cdot \ln\left(\theta(\tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_i)\right) + (1 - y_i) \cdot \ln\left(\theta(-\tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_i)\right)}$$

The cross-entropy error function

$$\boxed{E(\tilde{\mathbf{w}}) = -\ln\left(L(\tilde{\mathbf{w}})\right) = \sum_{i=1}^{n} y_i \cdot \ln\left(\frac{1}{\theta(\tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_i)}\right) + (1 - y_i) \cdot \ln\left(\frac{1}{1 - \theta(\tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_i)}\right)}$$

$$\nabla(\tilde{\mathbf{w}}) = \frac{\partial}{\partial \tilde{\mathbf{w}}}\left\{\ln\left(L(\tilde{\mathbf{w}})\right)\right\} = \frac{\partial}{\partial \tilde{\mathbf{w}}}\left\{\sum_{i=1}^{n} y_i \cdot \ln(\theta(z_i)) + (1 - y_i) \cdot \ln(\theta(-z_i))\right\}$$

$$z_i = \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_i$$

$$\nabla(\tilde{\mathbf{w}}) = \sum_{i=1}^{n} y_i \cdot \theta(-z_i) \cdot \tilde{\mathbf{x}}_i - (1 - y_i) \cdot \theta(z_i) \cdot \tilde{\mathbf{x}}_i$$

$$= \sum_{i=1}^{n} \left(y_i - \theta(\tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_i)\right) \cdot \tilde{\mathbf{x}}_i$$

Example 24.2 (Logistic Regression). Figure 24.2(a) shows the output of logistic regression modeling on the Iris principal components data, where the independent attributes X_1 and X_2 represent the first two principal components, and the binary response variable Y represents the type of Iris flower; $Y = 1$ corresponds to `Iris-virginica`, whereas $Y = 0$ corresponds to the two other Iris types, namely `Iris-setosa` and `Iris-versicolor`.

The fitted logistic model is given as

$$\tilde{\mathbf{w}} = (w_0, w_1, w_2)^T = (-6.79, -5.07, -3.29)^T$$

$$P(Y=1|\tilde{\mathbf{x}}) = \theta(\tilde{\mathbf{w}}^T\tilde{\mathbf{x}}) = \frac{1}{1+\exp\{6.79+5.07\cdot x_1 + 3.29\cdot x_2\}}$$

Figure 24.2(a) plots $P(Y=1|\tilde{\mathbf{x}})$ for various values of $\tilde{\mathbf{x}}$.

Given $\tilde{\mathbf{x}}$, if $P(Y=1|\tilde{\mathbf{x}}) \geq 0.5$, then we predict $\hat{y} = 1$, otherwise we predict $\hat{y} = 0$. Figure 24.2(a) shows that five points (shown in dark gray) are misclassified. For example, for $\tilde{\mathbf{x}} = (1, -0.52, -1.19)^T$ we have:

$$P(Y=1|\tilde{\mathbf{x}}) = \theta(\tilde{\mathbf{w}}^T\tilde{\mathbf{x}}) = \theta(-0.24) = 0.44$$

$$P(Y=0|\tilde{\mathbf{x}}) = 1 - P(Y=1|\tilde{\mathbf{x}}) = 0.54$$

Thus, the predicted response for $\tilde{\mathbf{x}}$ is $\hat{y} = 0$, whereas the true class is $y = 1$.

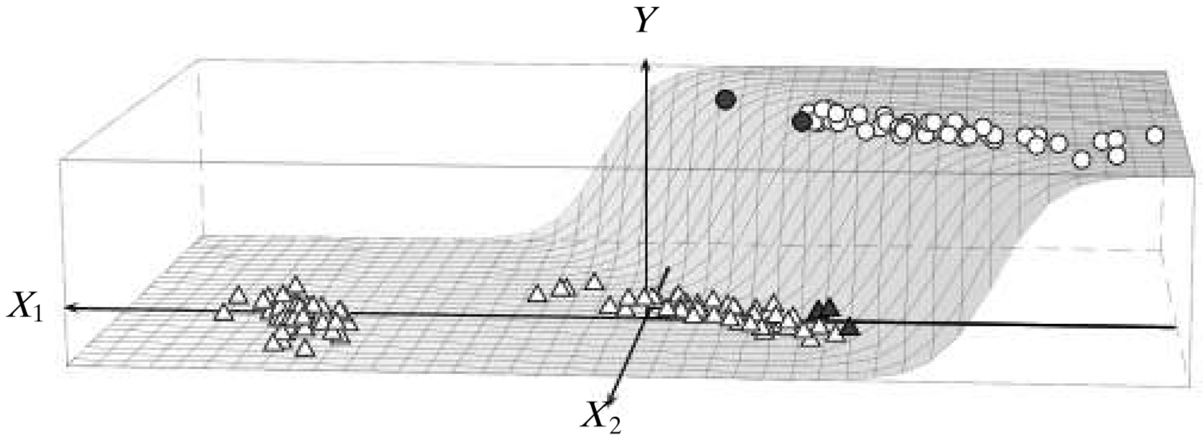
Figure 24.2 also contrasts logistic versus linear regression. The plane of best fit in linear regression is shown in Figure 24.2(b), with the weight vector:

$$\tilde{\mathbf{w}} = (0.333, -0.167, 0.074)^T$$

$$\hat{y} = f(\tilde{\mathbf{x}}) = 0.333 - 0.167\cdot x_1 + 0.074\cdot x_2$$

Since the response vector Y is binary, we predict the response class as $y = 1$ if $f(\tilde{\mathbf{x}}) \geq 0.5$, and $y = 0$ otherwise. The linear regression model results in 17 points being misclassified (dark gray points), as shown in Figure 24.2(b).

Since there are $n = 150$ points in total, this results in a training set or in-sample accuracy of 88.7% for linear regression. On the other hand, logistic regression misclassifies only 5 points, for an in-sample accuracy of 96.7%, which is a much better fit, as is also apparent in Figure 24.2.



(a) Logistic Regression

$$\hat{y} = \begin{cases} 1 & \text{if } \theta(\tilde{\mathbf{w}}^T\tilde{\mathbf{z}}) \geq 0.5 \\ 0 & \text{if } \theta(\tilde{\mathbf{w}}^T\tilde{\mathbf{z}}) < 0.5 \end{cases}$$

a.fahim@ut.ac.ir

$$Y \in \{c_1, c_2, \cdots, c_K\}.$$

ما Y را با یک متغیر تصادفی برنولی چند متغیره‌ی k بعدی مدل می‌کنیم.

از آنجایی‌که Y فقط یکی از k مقدار را می‌گیرد می‌توانیم از One-hot encoding استفاده کنیم.

$$\mathbf{Y} \in \{\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_K\} \qquad \mathbf{e}_i = (\overbrace{0, \ldots, 0}^{i-1}, 1, \overbrace{0, \ldots, 0}^{K-i})^T$$

$$P(\mathbf{Y} = \mathbf{e}_i | \tilde{\mathbf{X}} = \tilde{\mathbf{x}}) = \pi_i(\tilde{\mathbf{x}}), \text{ for } i = 1, 2, \ldots, K$$

$$\sum_{i=1}^{K} \pi_i(\tilde{\mathbf{x}}) = \sum_{i=1}^{K} P(\mathbf{Y} = \mathbf{e}_i | \tilde{\mathbf{X}} = \tilde{\mathbf{x}}) = 1$$

اگر $\boldsymbol{Y} = \boldsymbol{e}_i$ باشد آنگاه فقط $Y_i = 1$ و الباق Y_j ها صفر است.

$$\boxed{P(\mathbf{Y} | \tilde{\mathbf{X}} = \tilde{\mathbf{x}}) = \prod_{j=1}^{K} \left(\pi_j(\tilde{\mathbf{x}})\right)^{Y_j}}$$

در حالت چندکلاس باید یک کلاس را به عنوان مبنا انتخاب کنیم.

$$\pi_i(\tilde{\mathbf{x}}) = \exp\{\tilde{\boldsymbol{\omega}}_i^T \tilde{\mathbf{x}}\} \cdot \pi_K(\tilde{\mathbf{x}})$$

$$\ln(\text{odds}(\mathbf{Y} = \mathbf{e}_i | \tilde{\mathbf{X}} = \tilde{\mathbf{x}})) = \ln\left(\frac{P(\mathbf{Y} = \mathbf{e}_i | \tilde{\mathbf{X}} = \tilde{\mathbf{x}})}{P(\mathbf{Y} = \mathbf{e}_K | \tilde{\mathbf{X}} = \tilde{\mathbf{x}})}\right) = \ln\left(\frac{\pi_i(\tilde{\mathbf{x}})}{\pi_K(\tilde{\mathbf{x}})}\right) = \tilde{\boldsymbol{\omega}}_i^T \tilde{\mathbf{x}}$$

$$= \omega_{i0} \cdot x_0 + \omega_{i1} \cdot x_1 + \cdots + \omega_{id} \cdot x_d$$

$$\sum_{j=1}^{K} \pi_j(\tilde{\mathbf{x}}) = 1 \qquad \Longrightarrow \left(\sum_{j \neq K} \exp\{\tilde{\boldsymbol{\omega}}_j^T \tilde{\mathbf{x}}\} \cdot \pi_K(\tilde{\mathbf{x}})\right) + \pi_K(\tilde{\mathbf{x}}) = 1$$

$$\Longrightarrow \pi_K(\tilde{\mathbf{x}}) = \frac{1}{1 + \sum_{j \neq K} \exp\{\tilde{\boldsymbol{\omega}}_j^T \tilde{\mathbf{x}}\}}$$

$$\pi_i(\tilde{\mathbf{x}}) = \exp\{\tilde{\boldsymbol{\omega}}_i^T \tilde{\mathbf{x}}\} \cdot \pi_K(\tilde{\mathbf{x}}) = \frac{\exp\{\tilde{\boldsymbol{\omega}}_i^T \tilde{\mathbf{x}}\}}{1 + \sum_{j \neq K} \exp\{\tilde{\boldsymbol{\omega}}_j^T \tilde{\mathbf{x}}\}}$$

The Softmax function

$$\boxed{\pi_i(\tilde{\mathbf{x}}) = \frac{\exp\{\tilde{\boldsymbol{\omega}}_i^T \tilde{\mathbf{x}}\}}{\sum_{j=1}^{K} \exp\{\tilde{\boldsymbol{\omega}}_j^T \tilde{\mathbf{x}}\}}, \qquad \text{for all } i = 1, 2, \cdots, K}$$

برآورد بیشینه‌ی درستنمایی (Maximum Likelihood Estimation)

$$L(\tilde{\mathbf{W}}) = P(\mathbf{Y}|\tilde{\mathbf{W}}) = \prod_{i=1}^{n} P(\mathbf{y}_i|\tilde{\mathbf{X}} = \tilde{\mathbf{x}}_i) = \prod_{i=1}^{n}\prod_{j=1}^{K} \left(\pi_j(\tilde{\mathbf{x}}_i)\right)^{y_{ij}}$$

مجموعه‌ی K بردار وزن و K بردار پاسخ که One-hot encoding مدل شده است.

$$\mathbf{Y} \in \{\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_K\}. \qquad \tilde{\mathbf{W}} = \{\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2, \cdots, \tilde{\mathbf{w}}_K\}$$

$$\boxed{\ln\left(L(\tilde{\mathbf{W}})\right) = \sum_{i=1}^{n}\sum_{j=1}^{K} y_{ij} \cdot \ln(\pi_j(\tilde{\mathbf{x}}_i)) = \sum_{i=1}^{n}\sum_{j=1}^{K} y_{ij} \cdot \ln\left(\frac{\exp\{\tilde{\mathbf{w}}_j^T \tilde{\mathbf{x}}_i\}}{\sum_{a=1}^{K} \exp\{\tilde{\mathbf{w}}_a^T \tilde{\mathbf{x}}_i\}}\right)}$$

$$\frac{\partial}{\partial \pi_j(\tilde{\mathbf{x}}_i)} \ln(\pi_j(\tilde{\mathbf{x}}_i)) = \frac{1}{\pi_j(\tilde{\mathbf{x}}_i)}$$

$$\frac{\partial}{\partial \tilde{\mathbf{w}}_a} \pi_j(\tilde{\mathbf{x}}_i) = \begin{cases} \pi_a(\tilde{\mathbf{x}}_i) \cdot (1 - \pi_a(\tilde{\mathbf{x}}_i)) \cdot \tilde{\mathbf{x}}_i & \text{if } j = a \\ -\pi_a(\tilde{\mathbf{x}}_i) \cdot \pi_j(\tilde{\mathbf{x}}_i) \cdot \tilde{\mathbf{x}}_i & \text{if } j \neq a \end{cases}$$

$$\nabla(\tilde{\mathbf{w}}_a) = \frac{\partial}{\partial \tilde{\mathbf{w}}_a}\left\{\ln(L(\tilde{\mathbf{W}}))\right\} = \sum_{i=1}^{n} \left(y_{ia} - \pi_a(\tilde{\mathbf{x}}_i)\right) \cdot \tilde{\mathbf{x}}_i$$

$$\boxed{\nabla(\tilde{\mathbf{w}}_j, \tilde{\mathbf{x}}_i) = \left(y_{ij} - \pi_j(\tilde{\mathbf{x}}_i)\right) \cdot \tilde{\mathbf{x}}_i}$$

برای SGA وزن‌ها را در هر مرحله برای یک نقطه به روز می‌کنیم.

$$\boxed{\tilde{\mathbf{w}}_j^{t+1} = \tilde{\mathbf{w}}_j^t + \eta \cdot \nabla(\tilde{\mathbf{w}}_j^t, \tilde{\mathbf{x}}_i)}$$

$$\boxed{\hat{y} = \arg\max_{c_i}\{\pi_i(\tilde{\mathbf{z}})\} = \arg\max_{c_i}\left\{\frac{\exp\{\tilde{\mathbf{w}}_i^T \tilde{\mathbf{z}}\}}{\sum_{j=1}^{K} \exp\{\tilde{\mathbf{w}}_j^T \tilde{\mathbf{z}}\}}\right\}}$$

Example 24.3. Consider the Iris dataset, with $n = 150$ points in a 2D space spanned by the first two principal components, as shown in Figure 24.3. Here, the response variable takes on three values: $Y = c_1$ corresponds to `Iris-setosa` (shown as squares), $Y = c_2$ corresponds to `Iris-versicolor` (as circles) and $Y = c_3$ corresponds to `Iris-virginica` (as triangles). Thus, we map $Y = c_1$ to $\mathbf{e}_1 = (1,0,0)^T$, $Y = c_2$ to $\mathbf{e}_2 = (0,1,0)^T$ and $Y = c_3$ to $\mathbf{e}_3 = (0,0,1)^T$.

The multiclass logistic model uses $Y = c_3$ (`Iris-virginica`; triangles) as the reference or base class. The fitted model is given as:

$$\tilde{\mathbf{w}}_1 = (-3.52, 3.62, 2.61)^T$$
$$\tilde{\mathbf{w}}_2 = (-6.95, -5.18, -3.40)^T$$
$$\tilde{\mathbf{w}}_3 = (0, 0, 0)^T$$

Figure 24.3 plots the decision surfaces corresponding to the softmax functions:

$$\pi_1(\tilde{\mathbf{x}}) = \frac{\exp\{\tilde{\mathbf{w}}_1^T\tilde{\mathbf{x}}\}}{1 + \exp\{\tilde{\mathbf{w}}_1^T\tilde{\mathbf{x}}\} + \exp\{\tilde{\mathbf{w}}_2^T\tilde{\mathbf{x}}\}}$$
$$\pi_2(\tilde{\mathbf{x}}) = \frac{\exp\{\tilde{\mathbf{w}}_2^T\tilde{\mathbf{x}}\}}{1 + \exp\{\tilde{\mathbf{w}}_1^T\tilde{\mathbf{x}}\} + \exp\{\tilde{\mathbf{w}}_2^T\tilde{\mathbf{x}}\}}$$
$$\pi_3(\tilde{\mathbf{x}}) = \frac{1}{1 + \exp\{\tilde{\mathbf{w}}_1^T\tilde{\mathbf{x}}\} + \exp\{\tilde{\mathbf{w}}_2^T\tilde{\mathbf{x}}\}}$$

The surfaces indicate regions where one class dominates over the others. It is important to note that the points for c_1 and c_2 are shown displaced along Y to emphasize the contrast with c_3, which is the reference class.

Overall, the training set accuracy for the multiclass logistic classifier is 96.7%, since it misclassifies only five points (shown in dark gray). For example, for the point $\tilde{\mathbf{x}} = (1, -0.52, -1.19)^T$, we have:

$$\pi_1(\tilde{\mathbf{x}}) = 0 \qquad \pi_2(\tilde{\mathbf{x}}) = 0.448 \qquad \pi_3(\tilde{\mathbf{x}}) = 0.552$$

Thus, the predicted class is $\hat{y} = \arg\max_{c_i}\{\pi_i(\tilde{\mathbf{x}})\} = c_3$, whereas the true class is $y = c_2$.

Algorithm 24.2: Multiclass Logistic Regression Algorithm

LogisticRegression-MultiClass $(\mathbf{D}, \eta, \epsilon)$:

foreach $(\mathbf{x}_i^T, y_i) \in \mathbf{D}$ **do**
- $\tilde{\mathbf{x}}_i^T \leftarrow \begin{pmatrix} 1 & \mathbf{x}_i^T \end{pmatrix}$ // map to $\mathbb{R}^{d+1}$
- $\mathbf{y}_i \leftarrow \mathbf{e}_j$ if $y_i = c_j$ // map y_i to K-dimensional Bernoulli vector

$t \leftarrow 0$ // step/iteration counter

foreach $j = 1, 2, \cdots, K$ **do**
- $\tilde{\mathbf{w}}_j^t \leftarrow (0, \ldots, 0)^T \in \mathbb{R}^{d+1}$ // initial weight vector

repeat
- **foreach** $j = 1, 2, \cdots, K-1$ **do**
  - $\tilde{\mathbf{w}}_j \leftarrow \tilde{\mathbf{w}}_j^t$ // make a copy of $\tilde{\mathbf{w}}_j^t$
- **foreach** $\tilde{\mathbf{x}}_i \in \tilde{\mathbf{D}}$ *in random order* **do**
  - **foreach** $j = 1, 2, \cdots, K-1$ **do**
    - $\pi_j(\tilde{\mathbf{x}}_i) \leftarrow \dfrac{\exp\left\{\tilde{\mathbf{w}}_j^T\tilde{\mathbf{x}}_i\right\}}{\sum_{a=1}^K \exp\left\{\tilde{\mathbf{w}}_a^T\tilde{\mathbf{x}}_i\right\}}$
    - $\nabla(\tilde{\mathbf{w}}_j, \tilde{\mathbf{x}}_i) \leftarrow \left(y_{ij} - \pi_j(\tilde{\mathbf{x}}_i)\right) \cdot \tilde{\mathbf{x}}_i$ // compute gradient at $\tilde{\mathbf{w}}_j$
    - $\tilde{\mathbf{w}}_j \leftarrow \tilde{\mathbf{w}}_j + \eta \cdot \nabla(\tilde{\mathbf{w}}_j, \tilde{\mathbf{x}}_i)$ // update estimate for $\tilde{\mathbf{w}}_j$
- **foreach** $j = 1, 2, \cdots, K-1$ **do**
  - $\tilde{\mathbf{w}}_j^{t+1} \leftarrow \tilde{\mathbf{w}}_j$ // update $\tilde{\mathbf{w}}_j^{t+1}$
- $t \leftarrow t + 1$

until $\sum_{j=1}^{K-1} \|\tilde{\mathbf{w}}_j^t - \tilde{\mathbf{w}}_j^{t-1}\| \le \epsilon$

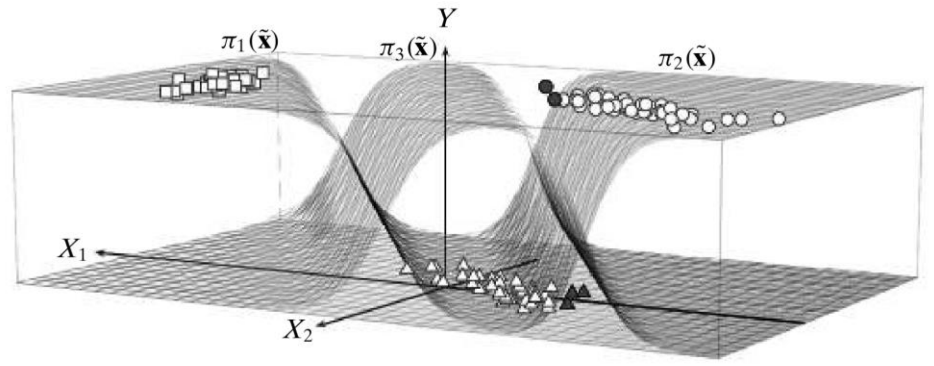


Figure 24.3. Multiclass logistic regression: Iris principal components data. Misclassified point are shown in dark gray color. All the points actually lie in the (X_1, X_2) plane, but c_1 and c_2 are shown displaced along Y with respect to the base class c_3 purely for illustration purposes.