Data 401
Logistic Regression

Dennis Sun

October 19, 2016

1. The Class So Far
2. Logistic Regression
3. Generalized Linear Models
4. Multinomial Logistic Regression

1. The Class So Far
2. Logistic Regression
3. Generalized Linear Models
4. Multinomial Logistic Regression

The Class So Far

The Class So Far

- We've been dealing with linear regression, which finds $\boldsymbol{\beta}$ that minimizes

$$\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}))^2.$$

The Class So Far

- We've been dealing with linear regression, which finds $\boldsymbol{\beta}$ that minimizes

$$\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}))^2.$$

- We saw last class that this is only appropriate if the response Y can be modeled as a linear function of the predictors, plus random normal errors.

The Class So Far

- We've been dealing with linear regression, which finds $\boldsymbol{\beta}$ that minimizes

$$\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}))^2.$$

- We saw last class that this is only appropriate if the response Y can be modeled as a linear function of the predictors, plus random normal errors.
- If the response is binary (i.e., classification problem), this assumption is preposterous!

1 The Class So Far

2 Logistic Regression

3 Generalized Linear Models

4 Multinomial Logistic Regression

Logistic Regression

Logistic Regression

- Suppose we still want a trend that's linear in our predictors.

$$\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

Logistic Regression

- Suppose we still want a trend that's linear in our predictors.

$$\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$$

- In linear regression, this models the mean μ_i in $Y_i \sim N(\mu_i, \sigma^2)$.

Logistic Regression

- Suppose we still want a trend that's linear in our predictors.

$$\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$$

- In linear regression, this models the mean μ_i in $Y_i \sim N(\mu_i, \sigma^2)$.
- In logistic regression, the responses are binary, so $Y_i \sim \text{Binom}(1, p_i)$. What should be modeled by the linear predictor?

Logistic Regression

- Suppose we still want a trend that's linear in our predictors.

$$\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$$

- In linear regression, this models the mean μ_i in $Y_i \sim N(\mu_i, \sigma^2)$.
- In logistic regression, the responses are binary, so $Y_i \sim \text{Binom}(1, p_i)$. What should be modeled by the linear predictor?
- It doesn't make sense to model p_i directly because $p_i \in [0, 1]$.

Logistic Regression

- Suppose we still want a trend that's linear in our predictors.

$$\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$$

- In linear regression, this models the mean μ_i in $Y_i \sim N(\mu_i, \sigma^2)$.
- In logistic regression, the responses are binary, so $Y_i \sim \text{Binom}(1, p_i)$. What should be modeled by the linear predictor?
- It doesn't make sense to model p_i directly because $p_i \in [0, 1]$.
- So we instead use the linear predictor to model the log-odds:

$$\log \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}.$$

Logistic Regression

- Suppose we still want a trend that's linear in our predictors.

$$\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$$

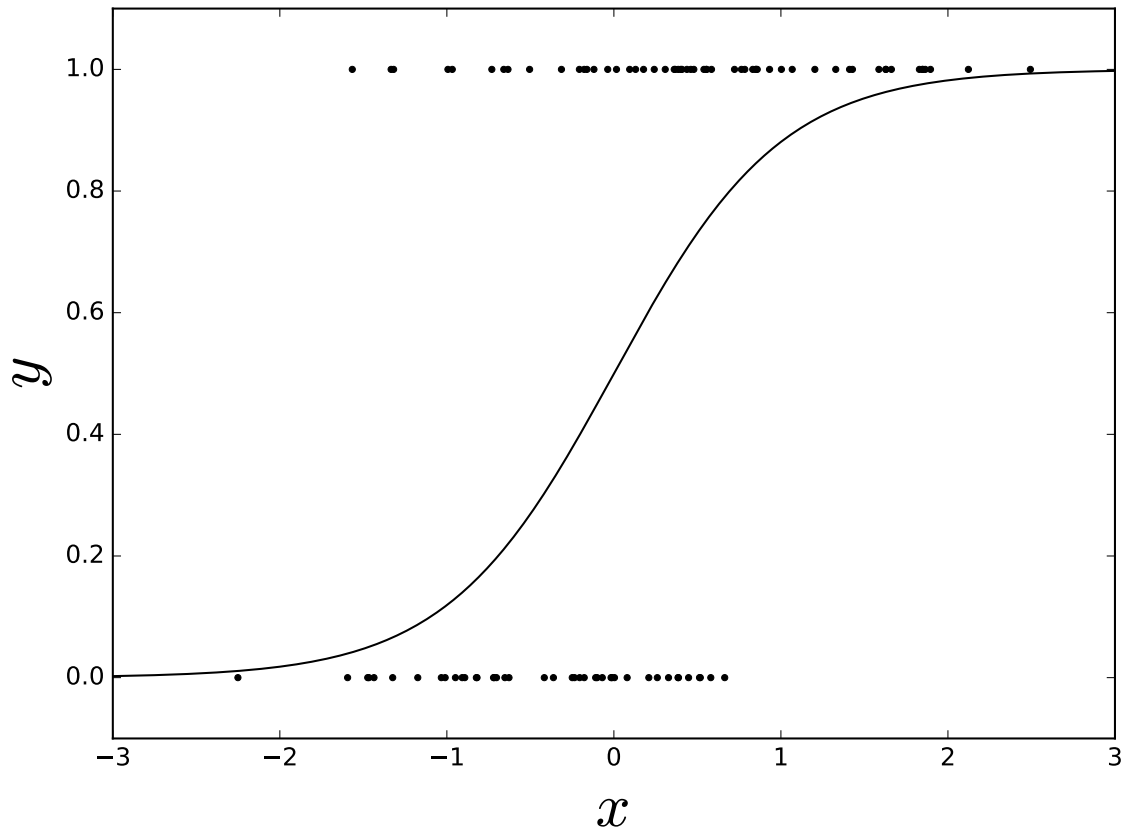
- In linear regression, this models the mean μ_i in $Y_i \sim N(\mu_i, \sigma^2)$.
- In logistic regression, the responses are binary, so $Y_i \sim \text{Binom}(1, p_i)$. What should be modeled by the linear predictor?
- It doesn't make sense to model p_i directly because $p_i \in [0, 1]$.
- So we instead use the linear predictor to model the log-odds:

$$\log \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}.$$

- This is equivalent to modeling p_i as:

$$p_i = \frac{e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}}}$$

Logistic Curve



Estimating Logistic Regression Coefficients

To estimate $\boldsymbol{\beta}$ in logistic regression, we use maximum likelihood:

Estimating Logistic Regression Coefficients

To estimate $\boldsymbol{\beta}$ in logistic regression, we use maximum likelihood:

$$p_{\boldsymbol{\beta}}(Y_1, ..., Y_n)$$

Estimating Logistic Regression Coefficients

To estimate $\boldsymbol{\beta}$ in logistic regression, we use maximum likelihood:

$$p_{\boldsymbol{\beta}}(Y_1, ..., Y_n) = \prod_{i=1}^{n} p_{\boldsymbol{\beta}}(Y_i)$$

Estimating Logistic Regression Coefficients

To estimate $\boldsymbol{\beta}$ in logistic regression, we use maximum likelihood:

$$
\begin{aligned}
p_{\boldsymbol{\beta}}(Y_1, ..., Y_n) &= \prod_{i=1}^{n} p_{\boldsymbol{\beta}}(Y_i) \\
&= \prod_{i=1}^{n} p_i(\boldsymbol{\beta})^{Y_i} (1 - p_i(\boldsymbol{\beta}))^{1-Y_i}
\end{aligned}
$$

Estimating Logistic Regression Coefficients

To estimate $\boldsymbol{\beta}$ in logistic regression, we use maximum likelihood:

$$\begin{aligned} p_{\boldsymbol{\beta}}(Y_1, ..., Y_n) &= \prod_{i=1}^{n} p_{\boldsymbol{\beta}}(Y_i) \\ &= \prod_{i=1}^{n} p_i(\boldsymbol{\beta})^{Y_i} (1 - p_i(\boldsymbol{\beta}))^{1-Y_i} \end{aligned}$$

where $p_i(\boldsymbol{\beta}) = \dfrac{e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}}}$. Note that

$1 - p_i(\boldsymbol{\beta}) = \dfrac{1}{1 + e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}}}$.

Estimating Logistic Regression Coefficients

To estimate $\boldsymbol{\beta}$ in logistic regression, we use maximum likelihood:

$$p_{\boldsymbol{\beta}}(Y_1, ..., Y_n) = \prod_{i=1}^{n} p_{\boldsymbol{\beta}}(Y_i)$$

$$= \prod_{i=1}^{n} p_i(\boldsymbol{\beta})^{Y_i} (1 - p_i(\boldsymbol{\beta}))^{1-Y_i}$$

where $p_i(\boldsymbol{\beta}) = \dfrac{e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}}}$. Note that

$1 - p_i(\boldsymbol{\beta}) = \dfrac{1}{1 + e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}}}$.

$$= \prod_{i=1}^{n} \left(\frac{p_i(\boldsymbol{\beta})}{1 - p_i(\boldsymbol{\beta})} \right)^{Y_i} (1 - p_i(\boldsymbol{\beta}))$$

Estimating Logistic Regression Coefficients

$$p_{\boldsymbol{\beta}}(Y_1, ..., Y_n) = \prod_{i=1}^{n} \left(\frac{p_i(\boldsymbol{\beta})}{1 - p_i(\boldsymbol{\beta})} \right)^{Y_i} (1 - p_i(\boldsymbol{\beta}))$$

Now let's take logs:

Estimating Logistic Regression Coefficients

$$p_{\boldsymbol{\beta}}(Y_1, ..., Y_n) = \prod_{i=1}^{n} \left(\frac{p_i(\boldsymbol{\beta})}{1 - p_i(\boldsymbol{\beta})} \right)^{Y_i} (1 - p_i(\boldsymbol{\beta}))$$

Now let's take logs:

$$\log p_{\boldsymbol{\beta}}(Y_1, ..., Y_n) = \sum_{i=1}^{n} Y_i \log \left(\frac{p_i(\boldsymbol{\beta})}{1 - p_i(\boldsymbol{\beta})} \right) + \sum_{i=1}^{n} \log(1 - p_i(\boldsymbol{\beta}))$$

Estimating Logistic Regression Coefficients

$$p_{\boldsymbol{\beta}}(Y_1, ..., Y_n) = \prod_{i=1}^{n} \left(\frac{p_i(\boldsymbol{\beta})}{1 - p_i(\boldsymbol{\beta})} \right)^{Y_i} (1 - p_i(\boldsymbol{\beta}))$$

Now let's take logs:

$$\begin{aligned} \log p_{\boldsymbol{\beta}}(Y_1, ..., Y_n) &= \sum_{i=1}^{n} Y_i \log \left(\frac{p_i(\boldsymbol{\beta})}{1 - p_i(\boldsymbol{\beta})} \right) + \sum_{i=1}^{n} \log(1 - p_i(\boldsymbol{\beta})) \\ &= \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}) \\ &\quad + \sum_{i=1}^{n} \log \frac{1}{1 + e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}}} \end{aligned}$$

Estimating Logistic Regression Coefficients

$$p_{\boldsymbol{\beta}}(Y_1, ..., Y_n) = \prod_{i=1}^{n} \left(\frac{p_i(\boldsymbol{\beta})}{1 - p_i(\boldsymbol{\beta})} \right)^{Y_i} (1 - p_i(\boldsymbol{\beta}))$$

Now let's take logs:

$$\begin{aligned} \log p_{\boldsymbol{\beta}}(Y_1, ..., Y_n) &= \sum_{i=1}^{n} Y_i \log \left(\frac{p_i(\boldsymbol{\beta})}{1 - p_i(\boldsymbol{\beta})} \right) + \sum_{i=1}^{n} \log(1 - p_i(\boldsymbol{\beta})) \\ &= \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}) \\ &\quad - \sum_{i=1}^{n} \log(1 + e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}}) \end{aligned}$$

Solving for Logistic Regression Coefficients

We need to find $\boldsymbol{\beta}$ that maximizes

$$\ell_{\mathbf{Y}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i(\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}) - \sum_{i=1}^{n} \log(1+e^{\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}}).$$

Solving for Logistic Regression Coefficients

We need to find $\boldsymbol{\beta}$ that maximizes

$$\ell_{\mathbf{Y}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i(\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}) - \sum_{i=1}^{n} \log(1+e^{\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}}).$$

- Traditional approach: Use Newton-Raphson (which requires second-derivatives, i.e., the Hessian).

Solving for Logistic Regression Coefficients

We need to find $\boldsymbol{\beta}$ that maximizes

$$\ell_{\mathbf{Y}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i(\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}) - \sum_{i=1}^{n} \log(1+e^{\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}}).$$

- Traditional approach: Use Newton-Raphson (which requires second-derivatives, i.e., the Hessian).

  Each Newton update can be cleverly rewritten as the solution to a weighted least-squares problem. This algorithm is called **iteratively reweighted least squares** or **Fisher scoring**.

Solving for Logistic Regression Coefficients

We need to find $\boldsymbol{\beta}$ that maximizes

$$\ell_{\mathbf{Y}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i(\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}) - \sum_{i=1}^{n} \log(1+e^{\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}}).$$

- Traditional approach: Use Newton-Raphson (which requires second-derivatives, i.e., the Hessian).

  Each Newton update can be cleverly rewritten as the solution to a weighted least-squares problem. This algorithm is called **iteratively reweighted least squares** or **Fisher scoring**.
- Of course, we can always use gradient descent.

Gradient Descent for Logistic Regression

$$\ell_{\mathbf{Y}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i(\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}) - \sum_{i=1}^{n} \log(1+e^{\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}}).$$

What is the gradient?

Gradient Descent for Logistic Regression

$$\ell_{\mathbf{Y}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i(\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}) - \sum_{i=1}^{n} \log(1+e^{\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}}).$$

What is the gradient?

$$\nabla \ell_{\mathbf{Y}}(\boldsymbol{\beta}) = X^T\mathbf{Y} -$$

Gradient Descent for Logistic Regression

$$\ell_{\mathbf{Y}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i(\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}) - \sum_{i=1}^{n} \log(1+e^{\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}}).$$

What is the gradient?

$$\nabla \ell_{\mathbf{Y}}(\boldsymbol{\beta}) = X^T\mathbf{Y} - X^T\mathbf{p}(\boldsymbol{\beta})$$

Gradient Descent for Logistic Regression

$$\ell_{\mathbf{Y}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i(\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}) - \sum_{i=1}^{n} \log(1+e^{\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}}).$$

What is the gradient?

$$\nabla \ell_{\mathbf{Y}}(\boldsymbol{\beta}) = X^T\mathbf{Y} - X^T\mathbf{p}(\boldsymbol{\beta}) = X^T(\mathbf{Y} - \mathbf{p}(\boldsymbol{\beta})).$$

Gradient Descent for Logistic Regression

$$\ell_{\mathbf{Y}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i(\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}) - \sum_{i=1}^{n} \log(1+e^{\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}}).$$

What is the gradient?

$$\nabla \ell_{\mathbf{Y}}(\boldsymbol{\beta}) = X^T\mathbf{Y} - X^T\mathbf{p}(\boldsymbol{\beta}) = X^T(\mathbf{Y} - \mathbf{p}(\boldsymbol{\beta})).$$

So the updates are

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + t_k X^T(\mathbf{Y} - \mathbf{p}(\boldsymbol{\beta}^{(k)})).$$

Gradient Descent for Logistic Regression

$$\ell_{\mathbf{Y}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i(\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}) - \sum_{i=1}^{n} \log(1+e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}).$$

What is the gradient?

$$\nabla \ell_{\mathbf{Y}}(\boldsymbol{\beta}) = X^T\mathbf{Y} - X^T\mathbf{p}(\boldsymbol{\beta}) = X^T(\mathbf{Y} - \mathbf{p}(\boldsymbol{\beta})).$$

So the updates are

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + t_k X^T(\mathbf{Y} - \mathbf{p}(\boldsymbol{\beta}^{(k)})).$$

Interpretation: Move in the direction of observations for which we are currently underpredicting; move in the opposite direction as observations for which we are currently overpredicting.

Logistic Regression in Scikit-Learn

```
from sklearn.linear_model import LogisticRegression
```

In-Class Exercise

Open the notebook **`Logistic Regression Exercise.ipynb`** *and fit a logistic regression model to the spambase data. Try playing around with various penalty parameters (e.g.,* **`l1`***, a.k.a. lasso) and cross validation.*

1 The Class So Far

2 Logistic Regression

3 Generalized Linear Models

4 Multinomial Logistic Regression

Generalized Linear Models

Linear and logistic regression are examples of **generalized linear models** (GLMs), where we assume

$$Y_i \sim p_{\theta_i}$$

and we model $\theta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$ as a linear function of the predictors.

Generalized Linear Models

Linear and logistic regression are examples of **generalized linear models** (GLMs), where we assume

$$Y_i \sim p_{\theta_i}$$

and we model $\theta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$ as a linear function of the predictors.

For linear regression, $\theta_i = \mu_i$, the mean of each Y_i.

Generalized Linear Models

Linear and logistic regression are examples of **generalized linear models** (GLMs), where we assume

$$Y_i \sim p_{\theta_i}$$

and we model $\theta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$ as a linear function of the predictors.

For linear regression, $\theta_i = \mu_i$, the mean of each Y_i.

For logistic regression, $\theta_i = \log \frac{p_i}{1-p_i}$, the log-odds of $P(Y_i = 1)$.

Generalized Linear Models

Linear and logistic regression are examples of **generalized linear models** (GLMs), where we assume

$$Y_i \sim p_{\theta_i}$$

and we model $\theta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$ as a linear function of the predictors.

For linear regression, $\theta_i = \mu_i$, the mean of each Y_i.

For logistic regression, $\theta_i = \log \frac{p_i}{1-p_i}$, the log-odds of $P(Y_i = 1)$.

For Poisson regression, $\theta_i = \log \lambda_i$, the log of the mean.

Generalized Linear Models

Linear and logistic regression are examples of **generalized linear models** (GLMs), where we assume

$$Y_i \sim p_{\theta_i}$$

and we model $\theta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$ as a linear function of the predictors.

For linear regression, $\theta_i = \mu_i$, the mean of each Y_i.

For logistic regression, $\theta_i = \log \frac{p_i}{1-p_i}$, the log-odds of $P(Y_i = 1)$.

For Poisson regression, $\theta_i = \log \lambda_i$, the log of the mean.

STAT 418 is a class about these types of models.

Generalized Linear Models in scikit-learn

Generalized Linear Models in scikit-learn

Not currently supported. This is how you know the package wasn't written by statisticians.

1 The Class So Far

2 Logistic Regression

3 Generalized Linear Models

4 Multinomial Logistic Regression

Multinomial Logistic Regression

If there are more than 2 classes, then we can use the likelihood of a multinomial distribution.

Multinomial Logistic Regression

If there are more than 2 classes, then we can use the likelihood of a multinomial distribution.

Suppose for each observation, we have a binary vector $\mathbf{Y}_i = (Y_{i1}, ..., Y_{iK})$ indicating which class observation i falls into.

Multinomial Logistic Regression

If there are more than 2 classes, then we can use the likelihood of a multinomial distribution.

Suppose for each observation, we have a binary vector $\mathbf{Y}_i = (Y_{i1}, ..., Y_{iK})$ indicating which class observation i falls into. Then the likelihood is:

$$p_{\boldsymbol{\beta}}(\mathbf{Y}_1, ..., \mathbf{Y}_n) = \prod_{i=1}^{n} \prod_{j=1}^{K} p_{ij}(\boldsymbol{\beta}_j)^{Y_{ij}}.$$

Multinomial Logistic Regression

If there are more than 2 classes, then we can use the likelihood of a multinomial distribution.

Suppose for each observation, we have a binary vector $\mathbf{Y}_i = (Y_{i1}, ..., Y_{iK})$ indicating which class observation i falls into. Then the likelihood is:

$$p_{\boldsymbol{\beta}}(\mathbf{Y}_1, ..., \mathbf{Y}_n) = \prod_{i=1}^{n} \prod_{j=1}^{K} p_{ij}(\boldsymbol{\beta}_j)^{Y_{ij}}.$$

Note that each class has its own coefficients $\boldsymbol{\beta}_j$.

Multinomial Logistic Regression

If there are more than 2 classes, then we can use the likelihood of a multinomial distribution.

Suppose for each observation, we have a binary vector $\mathbf{Y}_i = (Y_{i1}, ..., Y_{iK})$ indicating which class observation i falls into. Then the likelihood is:

$$p_{\boldsymbol{\beta}}(\mathbf{Y}_1, ..., \mathbf{Y}_n) = \prod_{i=1}^{n} \prod_{j=1}^{K} p_{ij}(\boldsymbol{\beta}_j)^{Y_{ij}}.$$

Note that each class has its own coefficients $\boldsymbol{\beta}_j$.

But the coefficients are tied together by the fact that the probabilities $\mathbf{p}_i$ of the classes have to sum to 1.

Multinomial Logistic Regression in scikit-learn

Multinomial Logistic Regression in scikit-learn

Use option `multi_class='multinomial'` in `LogisticRegression`.

Multinomial Logistic Regression in scikit-learn

Use option `multi_class='multinomial'` in `LogisticRegression`.

By default, it does one-vs-rest, which means it fits K separate logistic regressions, one for each class. Each logistic regression gives an estimated probability for each class. But these probabilities do not necessarily have to add up to 1.