Logistic regression

Two flavors

Fabian Pedregosa

October 3, 2017

UC Berkeley

What are the most important methods data science?

1.	/stable/modules/generated/sklearn.linear_model.LogisticRegression.html	554,672	(4.26%)
2.	/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html	500,891	(3.85%)
3.	/stable/modules/generated/sklearn.svm.SVC.html	488,652	(3.75%)
4.	/stable/modules/generated/sklearn.linear_model.LinearRegression.html	385,655	(2.96%)
5.	/stable/modules/generated/sklearn.cluster.KMeans.html	380,696	(2.92%)
6.	/stable/modules/generated/sklearn.decomposition.PCA.html	372,332	(2.86%)
7.	/stable/modules/generated/sklearn.model_selection.train_test_split.html	323,041	(2.48%)
8.	/stable/modules/generated/sklearn.model_selection.GridSearchCV.html	308,502	(2.37%)
9.	/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html	280,697	(2.16%)
10.	/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html	240,033	(1.84%)
11.	/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html	227,029	(1.74%)

Logistic regression

Pillar of supervised learning. One of the most common methods

Two motivations

- As a probabilistic model.
- Mathematical optimization.

Probabilistic model

Probabilistic view

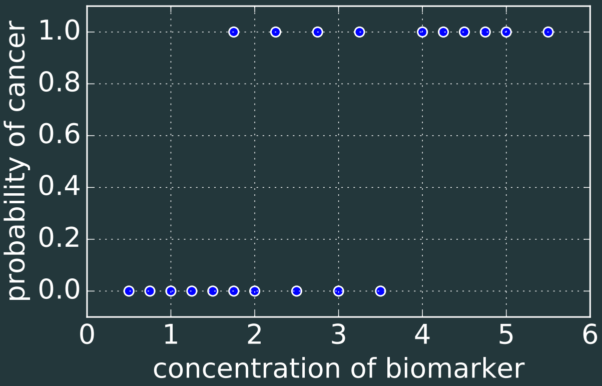
Motivation: classification problem with two classes.

Classes = "-1" and "1", which represent outcomes such as pass/fail, win/lose, alive/dead or healthy/sick, etc.

⚠ Despite its name, logistic regression is a model for classification and not regression.

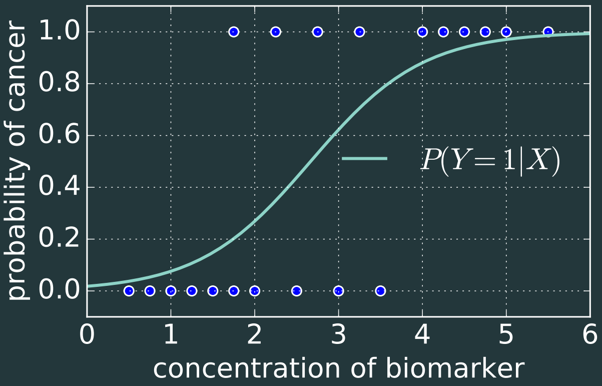
Probabilistic view

Motivation: Cancer / no cancer as a function of biomarker.



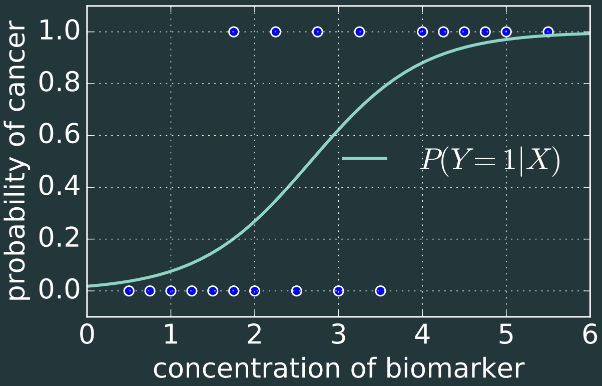
Probabilistic view

Motivation: Cancer / no cancer as a function of biomarker.



Probabilistic view

Motivation: Cancer / no cancer as a function of biomarker.



Goal: Given new data, estimate the probability of having cancer

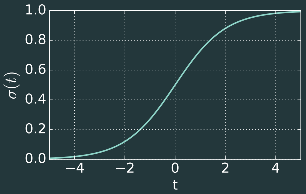
$\iff$ estimate $P(Y = 1|X)$

Probabilistic view

One popular model for $P(Y = 1|X)$ is the logistic model,

$$P(Y = y_i|X = x_i) = \sigma(y_i(x_i^T\beta_1 + \beta_0))$$

with $\sigma(t) = \dfrac{\exp(t)}{1 + \exp(t)}$



The logistic function $\sigma(t)$

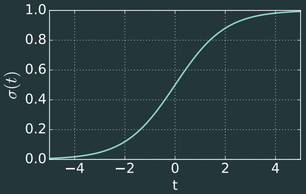
In the 1D case we have 2 degrees of freedom, β_1, β_0 that control the slope and intercept of the approximation.

Probabilistic view

One popular model for $P(Y = 1|X)$ is the logistic model,

$$P(Y = y_i|X = x_i) = \sigma(y_i(x_i^T \beta_1 + \beta_0))$$

$$\text{with } \sigma(t) = \frac{\exp(t)}{1 + \exp(t)}$$



The logistic function $\sigma(t)$

In the 1D case we have 2 degrees of freedom, β_1, β_0 that control the slope and intercept of the approximation. In the p-dimensional case, we have $p + 1$-degrees of freedom, as $\beta_1 \in \mathbb{R}^p, \beta_i \in \mathbb{R}$

Inference

The coefficients β_1, β_0 can be estimated as the ones that maximize the likelihood given the current data $\{(y_1, x_1), \ldots, (x_n, y_n)\}$.

$$\underset{\beta_1, \beta_0}{\text{maximize}} \; \ell(\beta_1, \beta_0) = \prod_{i=1}^{n} P(Y = y_i | X = x_i)$$

Inference

The coefficients β_1, β_0 can be estimated as the ones that maximize the likelihood given the current data $\{(y_1, x_1), \ldots, (x_n, y_n)\}$.

$$\underset{\beta_1, \beta_0}{\text{maximize}}\ \ell(\beta_1, \beta_0) = \prod_{i=1}^{n} P(Y = y_i | X = x_i)$$

Although for numerical reasons we rather minimize the minus log-likelihood

$$\begin{aligned} \underset{\beta_1, \beta_0}{\text{minimize}} - \log(\ell(\beta_1, \beta_0)) &= \sum_{i=1}^{n} - \log(P(Y = y_i | X = x_i)) && (1) \\ &= \sum_{i=1}^{n} \log(1 + \exp(-y_i(x_i^T \beta_1 + \beta_0))) && (2) \end{aligned}$$

Generalization

This approach can be naturally generalized to K classes.

$$\begin{aligned}
\Pr(Y_i = 1) &= \frac{e^{\boldsymbol{\beta}_1 \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol{\beta}_k \cdot \mathbf{X}_i}} \\
\Pr(Y_i = 2) &= \frac{e^{\boldsymbol{\beta}_2 \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol{\beta}_k \cdot \mathbf{X}_i}} && (3) \\
\cdots\cdots & && (4) \\
\Pr(Y_i = K-1) &= \frac{e^{\boldsymbol{\beta}_{K-1} \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol{\beta}_k \cdot \mathbf{X}_i}} && (5)
\end{aligned}$$

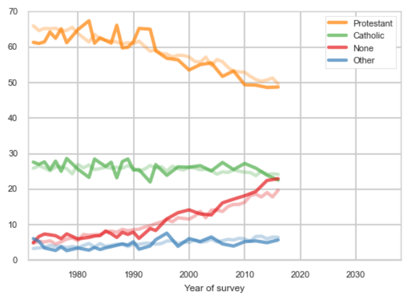
Application

Lets use this to predict the future!

Application

Lets use this to predict the future!

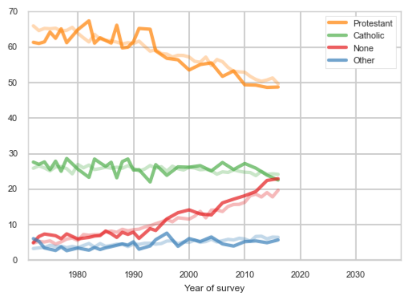
Use the data from the religion dataset to predict how religious beliefs will evolve after 2017



Application

Lets use this to predict the future!

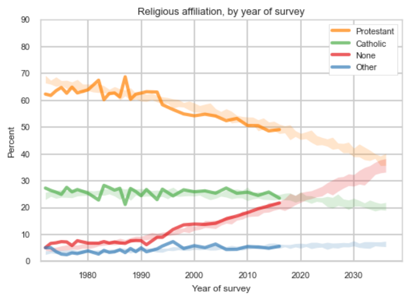
Use the data from the religion dataset to predict how religious beliefs will evolve after 2017



Suggestion: use scikit-learn's LogisticRegression class with multi_class='multinomial'. Use as features the year and cohort. The model can predict probabilities with method predict_proba.

Application

To predict, use the same data but with year and cohort shifted accordingly.

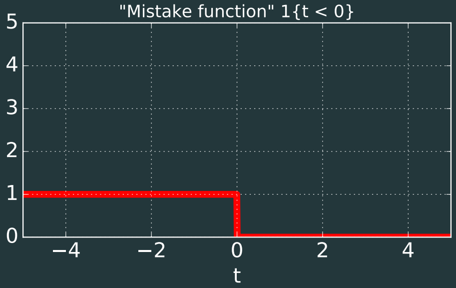


Optimization point of view

Optimization point of view

Setting. We have some data $\{(x_1, y_1), \ldots (x_n, y_n)\}, y_i \in \{-1, 1\}$, and we want to find the prediction rule $\hat{y}_i = \text{sign}(x_i^T \beta_1 + \beta_0)$ that makes less mistakes.

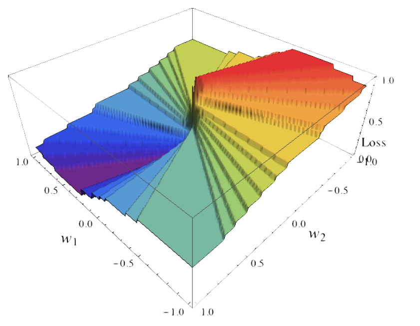
$$\underset{\beta_1, \beta_0}{\text{minimize}} \sum_{i=1}^{n} \mathbb{1}\{y_i(x_i^T \beta_1 + \beta_0) < 0\}$$



Optimization point of view

⚠Problems!

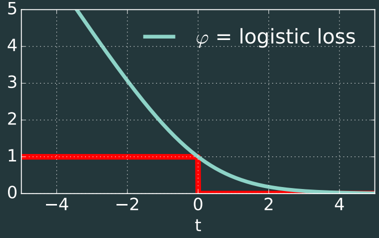
- Objective function is discontinuous
- Gradient is zero almost everywhere $\implies$ not amenable to gradient descent
- NP-hard problem in number of dimensions!



Optimization point of view

One way out: take a smooth upper bound on the "mistake function"

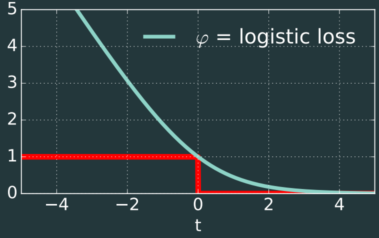
- $\varphi = \log(1 + \exp(-t))$
- Same function that appeared in minimization of log-likelihood (Eq (2)).



Optimization point of view

One way out: take a smooth upper bound on the "mistake function"

- $\varphi = \log(1 + \exp(-t))$
- Same function that appeared in minimization of log-likelihood (Eq (2)).



$\implies$ optimization becomes a convex and smooth problem.

Optimization point of view

Probabilistic approach

- Allows to interpret output as probabilities.

Optimization approach

- Makes links with other methods such as SVMs and neural networks.

References i

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe.
Convexity, classification, and risk bounds.
Journal of the American Statistical Association, 2006.

C. M. Bishop.
Pattern recognition and machine learning.
Springer, 2006.