Logistic Regression

Adam Richards

Galvanize, Inc

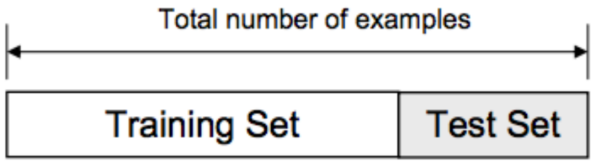
Last updated: 7. März 2018

Objectives

- The sigmoid function
- Logistic regression
- Validation / Confusion Matrix
- ROC Curve

Train test split

In practice we usually split our data into training data and test data. The training set contains a known output and the model learns on this data in order to be generalized to other data later on. We have the test dataset (or subset) in order to test how well our model generalized to unseen data.



https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6

Scaling with a train/test split

```
import numpy as np
from sklearn.preprocessing import StandardScaler
from sklearn.datasets import make_classification
from sklearn.model_selection import train_test_split

## create some data
X,y = make_classification(n_samples=50, n_features=5)

## make a train test split
X_train, X_test, y_train, y_test = train_test_split(X, y)

## scale using sklearn
scaler = StandardScaler().fit(X_train)
X_train_1 = scaler.transform(X_train)
X_test_1 = scaler.transform(X_test)

## scale without sklearn
X_train_2 = (X_train - X_train.mean(axis=0)) / X_train.std(axis=0)
X_test_2 = (X_test - X_train.mean(axis=0)) / X_train.std(axis=0)
```

Mean absolute error (MAE)

$$\sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{n} \tag{1}$$

```
from sklearn.metrics import mean_absolute_error
y_true = [3, -0.5, 2, 7]
y_pred = [2.5, 0.0, 2, 8]
mean_absolute_error(y_true, y_pred)
```

Root mean squared error (RMSE)

$$\sqrt{\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n}} \tag{2}$$

```
import numpy as np
from sklearn.metrics import mean_squared_error
y_true = [3, -0.5, 2, 7]
y_pred = [2.5, 0.0, 2, 8]
np.sqrt(mean_squared_error(y_true, y_pred))
```

Motivation for Logistic Regression

We are now moving into the world of classification problems. This is just like the regression problem, except that the values y we now want to predict take on only a small number of discrete values. For now, we will focus on the binary classification problem in which y can can be 0 and 1.

- benign/malignant, spam/ham, coffee/tea, pass/fail
- Most of what we describe here generalizes to the multi-class problem

What about linear regression?

We could approach the classification problem ignoring the fact that y is discrete-valued, and use our old linear regression algorithm to try to predict y given x. This does not always perform well.

Dogs and Horses

Does it make sense for our predicted values to take values larger than 1 or smaller than 0 when we know that $y \in 0, 1$?

To the Notebooks!

Dogs and Horses

Does it make sense for our predicted values to take values larger than 1 or smaller than 0 when we know that $y \in 0, 1$?

For this reason we use the following hypothesis

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}} \tag{3}$$

where,

$$g(z) = \frac{1}{1 + e^{-z}} \tag{4}$$

the parameters θ are also known as weights

Comparing linear and logistic regression

- In linear regression, the expected values of the target variable are modeled based on combination of values taken by the features
- In logistic regression the probability or odds of the target taking a particular value is modeled based on combination of values taken by the features.

Logistic function

The logistic function is also known as the sigmoid function.

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \text{ or } \frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X} \tag{5}$$

- We can think of probability as $p \sim \frac{\#successes}{\#trials}$
- We can think of the odds as $d = \frac{p}{1-p}$
- We can think of the log odds as $\theta = \ln(d) = \ln(\frac{p}{1-p})$
- $\theta = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$
- $\theta = \ln(\frac{p}{1-p})$
- $p = \frac{1}{1+e^{-\theta}}$

Objectives

- ✓ The sigmoid function
- Logistic regression
- Validation / Confusion Matrix
- ROC Curve

Logistic regression

Some perspective

Fisher proposed linear discriminant analysis in 1936. In the 1940s, various authors put forth an alternative approach, logistic regression. In the early 1970s, Nelder and Wedderburn coined the term generalized linear models for an entire class of statistical learning methods that include both linear and logistic regression as special cases. (Hastie et al., 2009) pp20.

Why might linear regression not be appropriate for the following?

- `y_label={1:'asthma',2:'lung cancer',3:'bronchitis'}`
- In logistic regression we are trying to model the probabilities of the K classes via linear functions in x
- These models are usually fit by MLE
- Rather than model the response directly (like in linear regression) logistic regression models the probability that Y belongs to a category
- e.g $P(\text{asthma} \mid \text{years\_smoked})$ is between 0 and 1 for any `years_smoked`

Optimization methods

Objective function

Any function for which we wish to find the minimum or maximum

In logistic regression the log-likelihood (prob. parameters given the data) for N observations can be specified as

$$\ell(\beta) = \sum_{i=1}^{N} \{y_i \log p(x_i; \beta) + (1 - y_i) \log(1 - p(x_i; \beta))\} \tag{6}$$

where $p(x; \beta)$ and $1 - p(x; \beta)$ are the probabilities of class 1 and class 2 in a $k = 2$ class scenario.
Recall that we wish to model $p(X)$ using the logistic function

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \text{ or } \frac{p(X)}{1 - p(X)} = e^{\beta_0 + \beta_1 X} \tag{7}$$

If $p(X) = 0.2$ then 1/5 people will have asthma with an odds of $\frac{0.2}{1-0.2} = \frac{1}{4}$.

(James et al., 2014) Chapter 4

Take the log of both sides of our logistic function then we get the logit or log-odds

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X \tag{8}$$

- How do we interpret β_1 in a linear regression setting?
- How do we interpret β_1 in a logistic regression setting?

We want to find $\hat{\beta}_0$ and $\hat{\beta}_1$ s.t. plugging in estimates for

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \tag{9}$$

close to 1 for individuals with asthma and close to 0 for those without

(James et al., 2014) Chapter 4

Take the log of both sides of our logistic function then we get the logit or log-odds

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X \tag{8}$$

- How do we interpret β_1 in a linear regression setting?
  β_1 gives the average change in Y associated with a one-unit increase in X
- How do we interpret β_1 in a logistic regression setting?

We want to find $\hat{\beta}_0$ and $\hat{\beta}_1$ s.t. plugging in estimates for

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \tag{9}$$

close to 1 for individuals with asthma and close to 0 for those without

(James et al., 2014) Chapter 4

Take the log of both sides of our logistic function then we get the logit or log-odds

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X \tag{8}$$

- How do we interpret β_1 in a linear regression setting?
  β_1 gives the average change in Y associated with a one-unit increase in X
- How do we interpret β_1 in a logistic regression setting?
  Increasing X by one unit changes the log odds by β_1

We want to find $\hat{\beta}_0$ and $\hat{\beta}_1$ s.t. plugging in estimates for

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \tag{9}$$

close to 1 for individuals with asthma and close to 0 for those without

(James et al., 2014) Chapter 4

Objectives

- ✓ The sigmoid function
- ✓ Logistic regression
  - Validation / Confusion Matrix
  - ROC Curve

In classification contexts, performance is assessed using a confusion matrix:

	Predicted False ($\hat{Y} = 0$)	Predicted True ($\hat{Y} = 1$)
Negative class ($Y = 0$)	True Negatives (TN)	False Positives (FP)
Positive class ($Y = 1$)	False Negatives(FN)	True Positives (TP)

There are many ways to evaluate the confusion matrix:

- Accuracy: overall proportion correct

$$\frac{TN + TP}{FP + FN + TN + TP}$$

- Precision: proportion called true that are correct

$$\frac{TP}{TP + FP}$$

- Recall: proportion of true that are called correctly

$$\frac{TP}{TP + FN}$$

- F_1-Score: balancing Precision/Recall

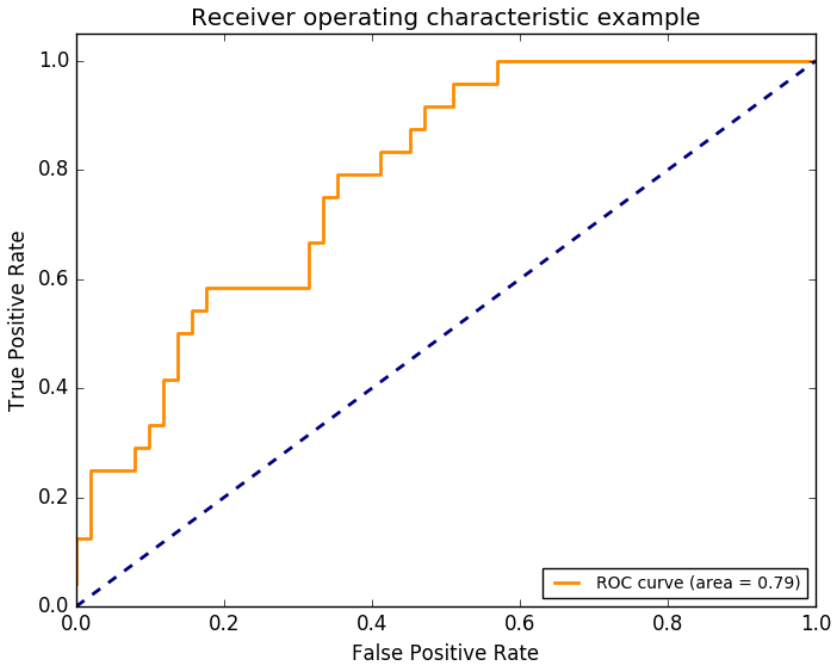
$$\frac{2}{\frac{1}{recall} + \frac{1}{precision}}$$

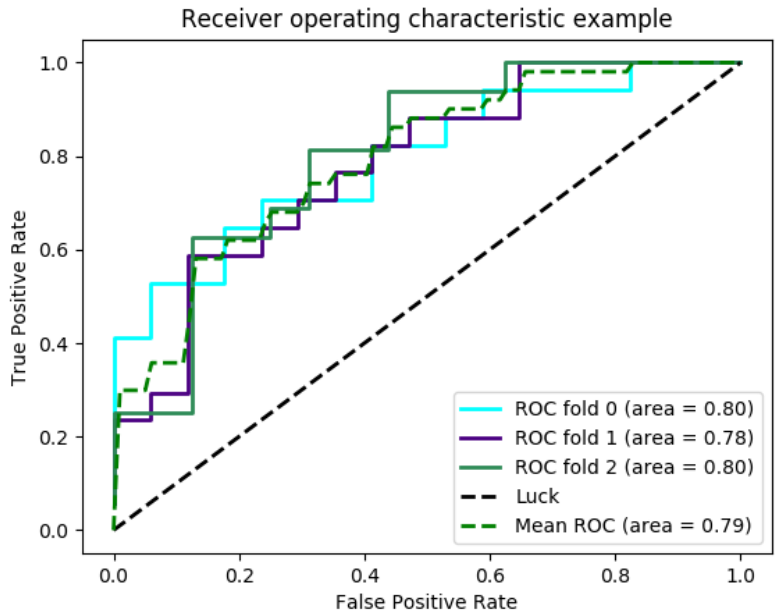
Exercise

Okay the last slide was very important break into groups of 3-4 and come up with a strategy to remember

1. How to fill out a confusion matrix
2. The formulas for: precision, recall, accuracy and F_1-Score

https://en.wikipedia.org/wiki/F1_score


Receiver operating characteristic example
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.79)


Receiver operating characteristic example
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC fold 0 (area = 0.80)
ROC fold 1 (area = 0.78)
ROC fold 2 (area = 0.80)
Luck
Mean ROC (area = 0.79)

Logistic Regression

```
import sklearn.linear_model as lm
help(lm.LogisticRegression)
```

Objectives

- ✓ The sigmoid function
- ✓ Logistic regression
- ✓ Validation / Confusion Matrix
- ✓ ROC Curve

References I

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2014). *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated.