# Logistic Regression

Adam Richards

Galvanize, Inc

Last updated: 28. September 2017

# Objectives

- Review
- The sigmoid function
- Logistic Regression
- Validation / Confusion Matrix
- ROC Curve

# Review

- Motivation - recall that the least squares approach to finding model parameters represents a specific case of maximum likelihood and overfitting is a general property of maximum likelihood estimation (MLE)
- So what again is regularization?

See pages 5-11 in (Bishop, 2006)

# Review

- Motivation - recall that the least squares approach to finding model parameters represents a specific case of maximum likelihood and overfitting is a general property of maximum likelihood estimation (MLE)
- So what again is regularization? - technique to control overfitting by introducing an a penalty term over the error function to discourage coefficients from reaching large values

$$\widetilde{E}(\mathbf{w}) = \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 + \lambda \|\mathbf{w}\|^2 \tag{1}$$

where

$$\|\mathbf{w}\|^2 = \boldsymbol{w}^T \boldsymbol{w} = w_0^2 + w_1^2 \dots w_n^2 \tag{2}$$

Note that $\lambda$ governs the relative importance of the regularization term compared with the SSE term

See pages 5-11 in (Bishop, 2006)

# More on shrinkage methods

- Why do we use the term shrinkage?
- Lasso Regression?
- Ridge Regression?

When we penalize by the sum of square errors in neural networks it is known as weight decay See pages 5-11 in (Bishop, 2006)

# More on shrinkage methods

- Why do we use the term shrinkage? Regularization is also referred to as shrinkage because it reduced the values of the coefficients
- Lasso Regression?
- Ridge Regression?

When we penalize by the sum of square errors in neural networks it is known as weight decay See pages 5-11 in (Bishop, 2006)

# More on shrinkage methods

- Why do we use the term shrinkage? Regularization is also referred to as shrinkage because it reduced the values of the coefficients
- Lasso Regression?

$$\hat{\mathbf{w}}^{\text{lasso}} = \underset{\mathbf{w}}{\operatorname{argmin}} \left\{ = \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 + \lambda \|\mathbf{w}\|_1 \right\} \tag{3}$$

where

$$\|w\|_1 = \sum_{j=1}^{M} |w_j|$$

- Ridge Regression?

When we penalize by the sum of square errors in neural networks it is known as weight decay See pages 5-11 in (Bishop, 2006)

# More on shrinkage methods

- Why do we use the term shrinkage? Regularization is also referred to as shrinkage because it reduced the values of the coefficients
- Lasso Regression?

$$\hat{\mathbf{w}}^{\text{lasso}} = \underset{\mathbf{w}}{\text{argmin}} \left\{ = \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 + \lambda \|\mathbf{w}\|_1 \right\} \tag{3}$$

where

$$\|w\|_1 = \sum_{j=1}^{M} |w_j|$$

- Ridge Regression?

$$\hat{\mathbf{w}}^{\text{ridge}} = \underset{\mathbf{w}}{\text{argmin}} \left\{ = \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 + \lambda \|\mathbf{w}\|_2^2 \right\} \tag{4}$$

$$\|w\|_2 = \sum_{j=1}^{M} w_j^2$$

When we penalize by the sum of square errors in neural networks it is known as weight decay See pages 5-11 in (Bishop, 2006)

# L1 and L2 penalties

**Interpretation**

When two predictors are highly correlated L1 penalties tend to pick one of the two while L2 will take both and shrink the coefficients

- In general L1 penalties are better at recovering sparse signals
- L2 penalties are better at minimizing prediction error

- So what type of regression is good for eliminating correlated variables?
- And if I just want to reduce the influence of two correlated variables?
- But what I just do not know which to use?

# Elastic net

The term elastic net refers to elastic net penalty to fit a generalized linear model (GLM)

Objective function is

loss + penalty

$$\min_{\beta_0,\beta} \frac{1}{N} \sum_{i=1}^{N} w_i l(y_i, \beta_0 | \beta^T x_i) \tag{5}$$

$$+ \lambda \left( (1-\alpha) \|\beta\|_2^2 /2 + \alpha \|\beta\|_1 \right) \tag{6}$$

where $\beta_0$ and $\beta$ are the coefficients of the GLM and $w$ is the weight of a given observation. The loss is $w$ times the negative log likelihood function and we see that $\alpha$ controls the balance between the type of penalty. $\lambda$ modulates the shrinkage.

(Hui and Hastie, 2005)

# Objectives

✓ Review

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

# Motivation

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

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}} \tag{7}$$

where,

$$g(z) = \frac{1}{1 + e^{-z}} \tag{8}$$

the parameters $\theta$ are also known as weights

# Terminology

We are doing Supervised learning: models using labels paired with features which can roughly be broken into:

- Regression: $y$ is continuous (price, demand, size)
- Classification: $y$ is categorical or discrete (fraud, churn)

| Machine-learning | Other fields |
|---|---|
| Features $X$ | Covariates, independent variables, regressors |
| Targets $y$ | dependent variable, regressand |
| Training | learning, estimation, model fitting |

**Logistic regression is classification?**

The output of a logistic regression model is (a transformation of) $(Y|X)$. So in a sense it is still regression.

# 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.

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \text{ or } \frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X} \quad (9)$$

- 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

- ✓ Review
- ✓ 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

$$\ell(\beta) = \sum_{i=1}^{N} \{y_i \log p(x_i; \beta) + (1 - y_i) \log(1 - p(x_i; \beta))\} \tag{10}$$

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

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \text{ or } \frac{p(X)}{1 - p(X)} = e^{\beta_0 + \beta_1 X} \tag{11}$$

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

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X \tag{12}$$

- How do we interpret $\beta_1$ in a linear regression setting?
- How do we interpret $\beta_1$ in a logistic regression setting?

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

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \tag{13}$$

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

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X \tag{12}$$

- How do we interpret $\beta_1$ in a linear regression setting?
  $\beta_1$ gives the average change in $Y$ associated with a one-unit increase in $X$
- How do we interpret $\beta_1$ in a logistic regression setting?

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

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \tag{13}$$

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

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X \tag{12}$$

- How do we interpret $\beta_1$ in a linear regression setting?
  $\beta_1$ gives the average change in $Y$ associated with a one-unit increase in $X$
- How do we interpret $\beta_1$ in a logistic regression setting?
  Increasing $X$ by one unit changes the log odds by $\beta_1$

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

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \tag{13}$$

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

(James et al., 2014) Chapter 4

# Objectives

- ✓ Review
- ✓ 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

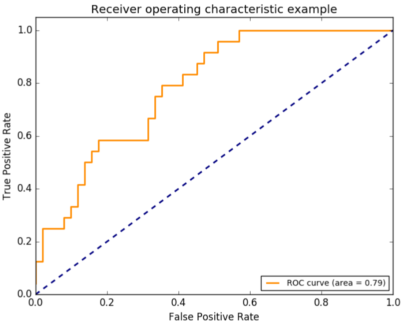
$$\frac{2}{\frac{1}{recall} + \frac{1}{precision}}$$

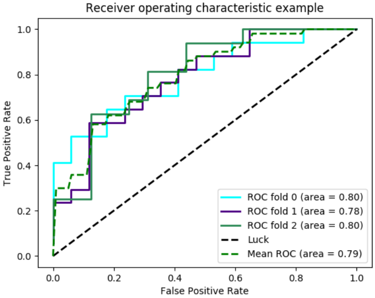
## Exercise

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

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

# References I

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

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

Hui, Z. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320.

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