Machine Learning in Business
John C. Hull

Chapter 3

Supervised Learning: Linear and Logistic Regression

Linear Regression

- Linear regression is a very popular tool because once you have made the assumption that the model is linear you do not need huge amount of data
- In ML we refer to the constant term as the bias and the coefficients as weights

Linear Regression *continued*

Assume n observations and m features. Model is

$$Y = a + b_1 X_1 + b_2 X_2 + \ldots . + b_m X_m + \varepsilon$$

Standard approach is to choose a and the b_i to minimize the mean square error (mse).

$$\text{mse} = \frac{1}{n} \sum_{j=1}^{n} \left[Y_j - \left(a + b_1 X_{1,j} + b_2 X_{2,j} + \ldots + b_m X_{m,j} \right) \right]^2$$

This can be done analytically by inverting a matrix. Alternatively a numerical (gradient descent) method can be used

Gradient Descent (brief description: more details in Chapter 6)

- The objective is to minimize a function by changing parameters. Steps are as follows:
  1. Choose starting value for parameters
  2. Find the steepest slope: i.e. the direction in which parameter have to be changed to reduce the objective function by the greatest amount
  3. Take a step down the valley in the direction of the steepest slope
  4. Repeat steps 2 and 3
  5. Continue until you reach the bottom of the valley

Categorical Features

- Categorical features are features where there are a number of non-numerical alternatives
- We can define a dummy variable for each alternative. The variable equals 1 if the alternative is true and zero otherwise. This is known as one-hot encoding
- But sometimes we do not have to do this because there is a natural ordering of variables, e.g.:

    small=1, medium=2, large=3

    assist. prof=1, assoc. prof=2, full prof =3

Dummy Variably Trap

- Suppose we have a constant term and a number of dummy variables (equal to 0 or 1)
- There is then no unique solution because, for any C, we can add C to the constant term and subtract C from each of the dummy variables without changing the prediction
- A side effect of regularization is that it solves this problem

Regularization

- Linear regression can over-fit, particularly when there are a large number of correlated features.
- Results for validation set may not then be as good as for training set
- Regularization is a way of avoiding overfitting and reducing the number of features. Alternatives:
  - Ridge
  - Lasso
  - Elastic net
- We must first scale feature values

Ridge regression (analytic solution)

- Reduce magnitude of regression coefficients by choosing a parameter λ and minimizing

$$\text{mse} + \lambda \sum_{i=1}^{m} b_i^2$$

- What happens as λ increases?

Lasso Regression (must use gradient descent)

- Similar to ridge regression except we minimize

$$\text{mse} + \lambda \sum_{i=1}^{m} |b_i|$$

- This has the effect of completely eliminating the less important factors

Elastic Net Regression *(must use gradient descent)*

- Middle ground between Ridge and Lasso
- Minimize

$$\text{mse} + \lambda_1 \sum_{i=1}^{m} b_i^2 + \lambda_2 \sum_{i=1}^{m} |b_i|$$

Baby Example (from Chapter 1)

Age (years)	Salary ($)
25	135,000
55	260,000
27	105,000
35	220,000
60	240,000
65	265,000
45	270,000
40	300,000
50	265,000
30	105,000

Baby Example *continued*

- We apply regularization to the model:

$$Y = a + b_1X + b_2X^2 + b_3X^3 + b_4X^4 + b_5X^5$$

where Y is salary and X is age

Data with Z-score scaling (Table 3.3)

Observ.	X	X^2	X^3	X^4	X^5
1	−1.290	−1.128	−0.988	−0.874	−0.782
2	0.836	0.778	0.693	0.592	0.486
3	−1.148	−1.046	−0.943	−0.850	−0.770
4	−0.581	−0.652	−0.684	−0.688	−0.672
5	1.191	1.235	1.247	1.230	1.191
6	1.545	1.731	1.901	2.048	2.174
7	0.128	−0.016	−0.146	−0.253	−0.333
8	−0.227	−0.354	−0.449	−0.511	−0.544
9	0.482	0.361	0.232	0.107	−0.004
10	−0.936	−0.910	−0.861	−0.803	−0.745

Ridge Results, Table 3.4 (λ=0.02 is similar to quadratic model)

λ	a	b_1	b_2	b_3	b_4	b_5
0	216.5	−32,623	135,403	−215,493	155,315	−42,559
0.02	216.5	97.8	36.6	−8.5	35.0	−44.6
0.10	216.5	56.5	28.1	3.7	−15.1	−28.4

Lasso Results, Table 3.5 (λ=1 is similar to the quadratic model)

λ	a	b_1	b_2	b_3	b_4	b_5
0	216.5	−32,623	135,403	−215,493	155,315	−42,559
0.02	216.5	−646.4	2,046.6	0.0	−3,351.0	2,007.9
0.1	216.5	355.4	0.0	−494.8	0.0	196.5
1	216.5	147.4	0.0	0.0	−99.3	0.0

Elastic Net Results: λ₁ = 0.02, λ₂=1

$$Y = 216.5 + 96.7X + 21.1X^2 - 26.0X^4 - 45.5X^5$$

Iowa House Price Case Study

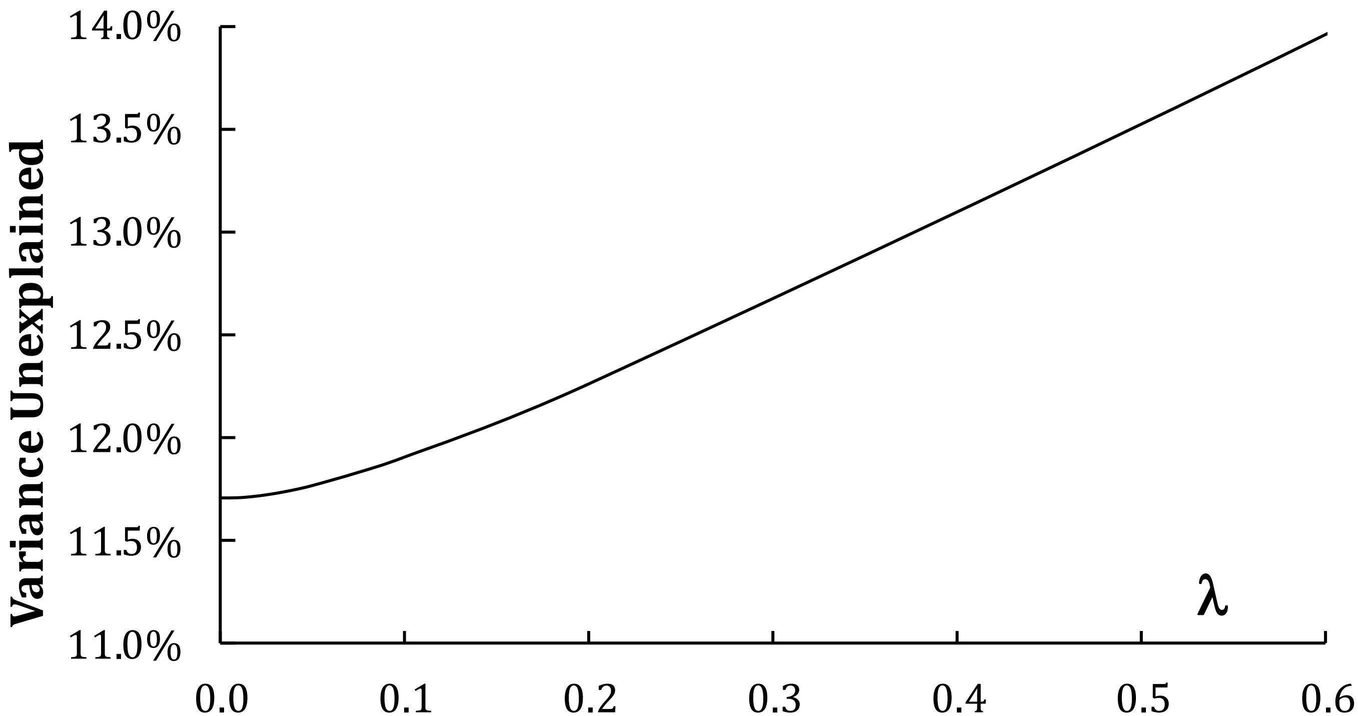
- The objective is to predict the prices of house in Iowa from features
- 800 observations in training set, 600 in validation set, and 508 in test set

Iowa House Price Results (No regularization)

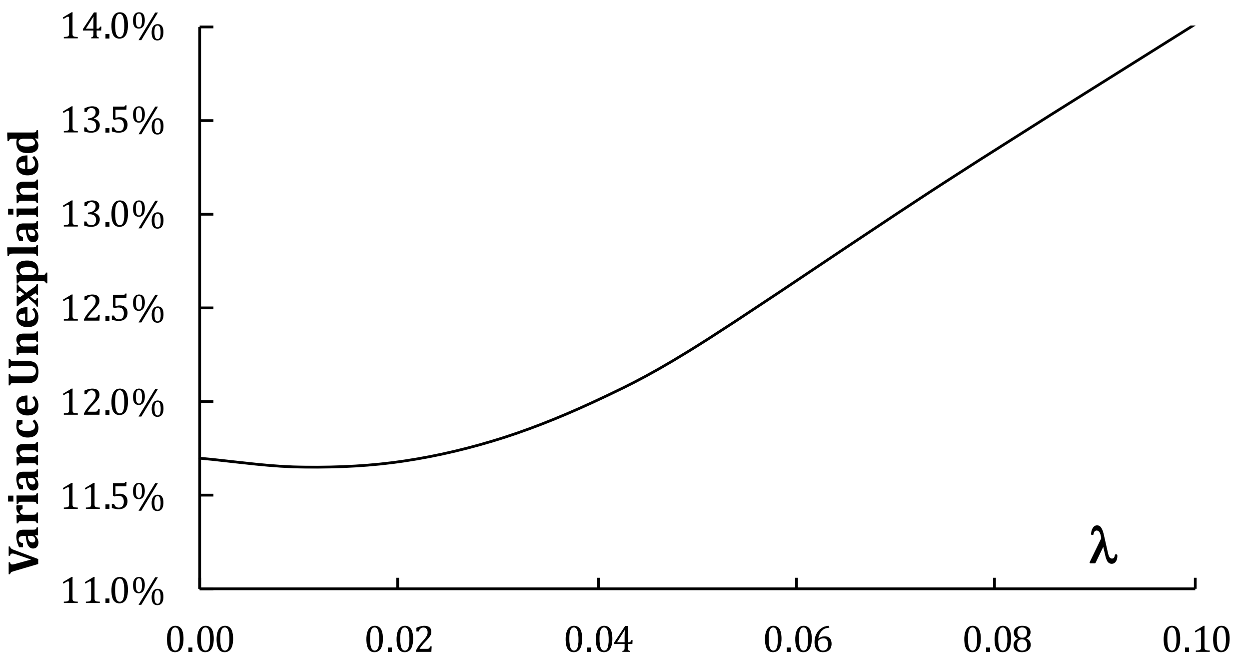
2 categorical variables included. Natural ordering for Basement quality. 25 dummy variables created for neighborhood

Lot area (squ ft)	0.08	Number of half bathrooms	0.02
Overall quality (scale from 1 to 10)	0.21	Number of bedrooms	−0.08
Overall condition (scale from 1 to 10)	0.10	Total rooms above grade	0.08
Year built	0.16	Number of fireplaces	0.03
Year remodeled	0.03	Parking spaces in garage	0.04
Basement finished squ ft	0.09	Garage area (squ ft)	0.05
Basement unfinished squ ft	−0.03	Wood deck (squ ft)	0.02
Total basement squ ft	0.14	Open porch (squ ft)	0.03
1st floor squ ft	0.15	Enclosed porch (squ ft0	0.01
2nd floor squ ft	0.13	Neighborhood (25 alternatives)	−0.05 to 0.12
Living area	0.16	Basement quality (6 natural ordering)	0.01
Number of full bathrooms	−0.02		

Ridge Results for validation set (Figure 3.8)



Lasso Results for validation set (Figure 3.9)



Non-zero weights for Lasso when λ=0.1 (overall quality and total living area were most important)

Feature	Weight
Lot Area (square feet)	0.04
Overall quality (Scale from 1 to 10)	0.30
Year built	0.05
Year remodeled	0.06
Finished basement (square feet)	0.12
Total basement (square feet)	0.10
First floor (square feet)	0.03
Living area (square feet)	0.30
Number of fireplaces	0.02
Parking spaces in garage	0.03
Garage area (square feet)	0.07
Neighborhoods (3 out of 25 non-zero)	0.01, 0.02, and 0.08
Basement quality	0.02

Summary of Iowa House Price Results

- With no regularization correlation between features leads to some negative weights which we would expect to be positive
- Improvements from Ridge is modest
- Lasso leads to a much bigger improvement in this case
- Elastic net similar to Lasso in this case
- Mean squared error for test set for Lasso with λ=0.1 is 14.7% so that 85.3% of variance is explained

Logistic Regression

- The objective is to classify observations into a "positive outcome" and "negative outcome" using data on features
- Probability of a positive outcome is assumed to be a sigmoid function:

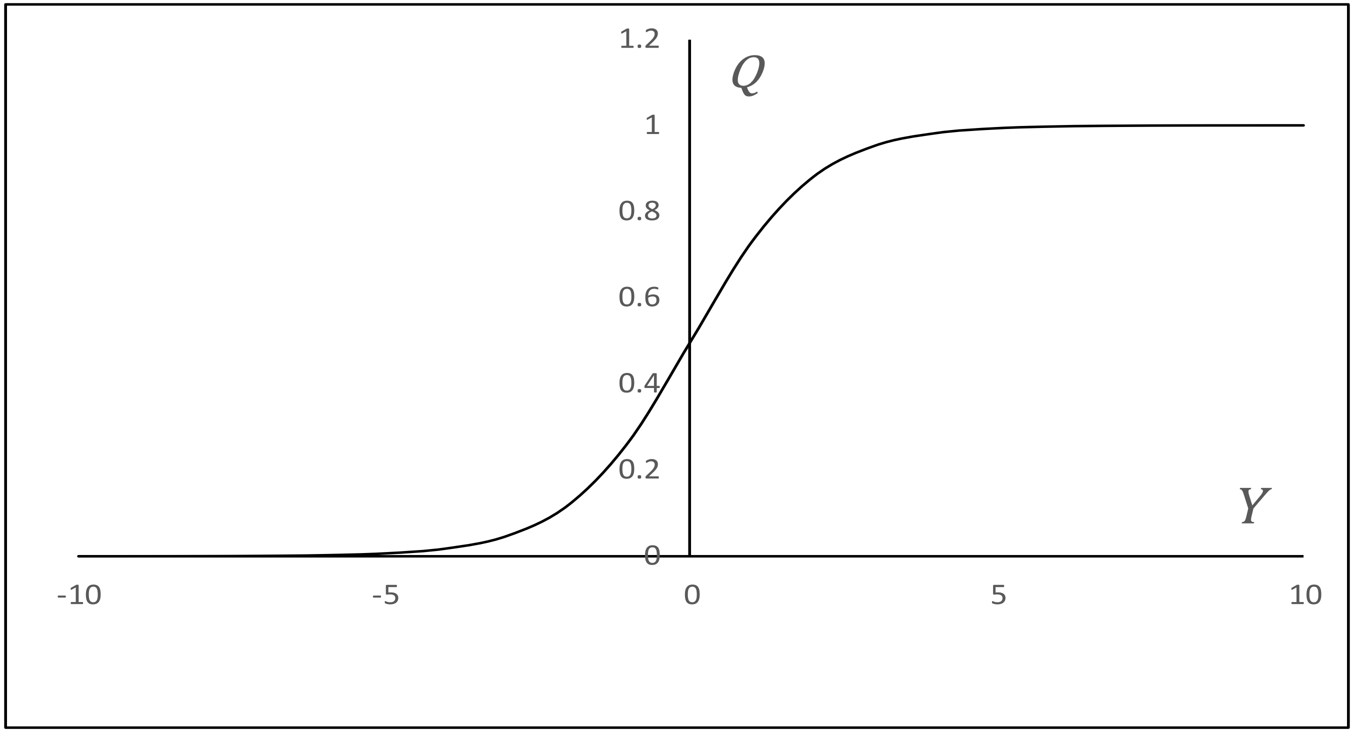
$$Q = \frac{1}{1 + e^{-Y}}$$

where Y is related linearly to the values of the features:

$$Y = a + b_1X_1 + b_2X_2 + \cdots + X_m$$

- Can use regularization

The Sigmoid Function (Figure 3.10)



Maximum Likelihood Estimation

- We use the training set to maximize

$$\sum_{\substack{\text{Positive}\\\text{Outcomes}}} \ln(Q) + \sum_{\substack{\text{Negative}\\\text{Outcomes}}} \ln(1-Q)$$

- This cannot be maximized analytically but we can use a gradient ascent algorithm

Lending Club Case Study

- Data consists of loans made and whether they proved to be good or defaulted. (A restriction is that you do not have data for loans that were never made.)
- We use only four features
  - Home ownership (rent vs. own)
  - Income
  - Debt to income
  - Credit score
- Training set has 8,695 observations (7,196 good loans and 1,499 defaulting loans). Test set has 5,196 observations (4,858 good loans and 1,058 defaulting loans)

The Data (Table 3.8)

Home Ownership 1=owns, 0 =rents	Income ($'000)	Debt to Income (%)	Credit score	1=Good, 0=Default
1	44.304	18.47	690	0
1	136.000	20.63	670	1
0	38.500	33.73	660	0
1	88.000	5.32	660	1
	….			….
	…..			….

Results for Lending Club Training Set

X_1= *Home Ownership*

X_2= *Income*

X_3= *Debt to income ratio*

X_4 = *Credit score*

$$Y = -6.5645 + 0.1395X_1 + 0.0041X_2 - 0.0011X_3 + 0.0113X_4$$

Decision Criterion

- The data set is imbalanced with more good loans than defaulting loans
- There are procedures for creating a balanced data set
- With a balanced data set we could classify an observation as positive if $Q > 0.5$ and negative otherwise
- However this does not consider the cost of misclassifying a bad loan and the lost profit from misclassifying a good loan
- A better approach is to investigate different thresholds, Z
  - If $Q > Z$ we accept a loan
  - If $Q \leq Z$ we reject the loan

Test Set Results (Tables 3.10, 3.11, and 3.12)

$Z = 0.75$:

	Predict no default	Predict default
Outcome positive (no default)	77.59%	4.53%
Outcome negative (default)	16.26%	1.62%

$Z=0.80$:

	Predict no default	Predict default
Outcome positive (no default)	55.34%	26.77%
Outcome negative (default)	9.75%	8.13%

$Z=0.85$:

	Predict no default	Predict default
Outcome positive (no default)	28.65%	53.47%
Outcome negative (default)	3.74%	14.15%

The Confusion matrix and common ratios

	Predict positive outcome	Predict negative outcome
Outcome positive	TP	FN
Outcome negative	FP	TN

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}}$$

$$\text{True Positive Rate (TPR also called sensitivity or recall)} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{The True Negative rate(also called specificity)} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{The False Positive Rate} = \frac{\text{FP}}{\text{TN} + \text{FP}}$$

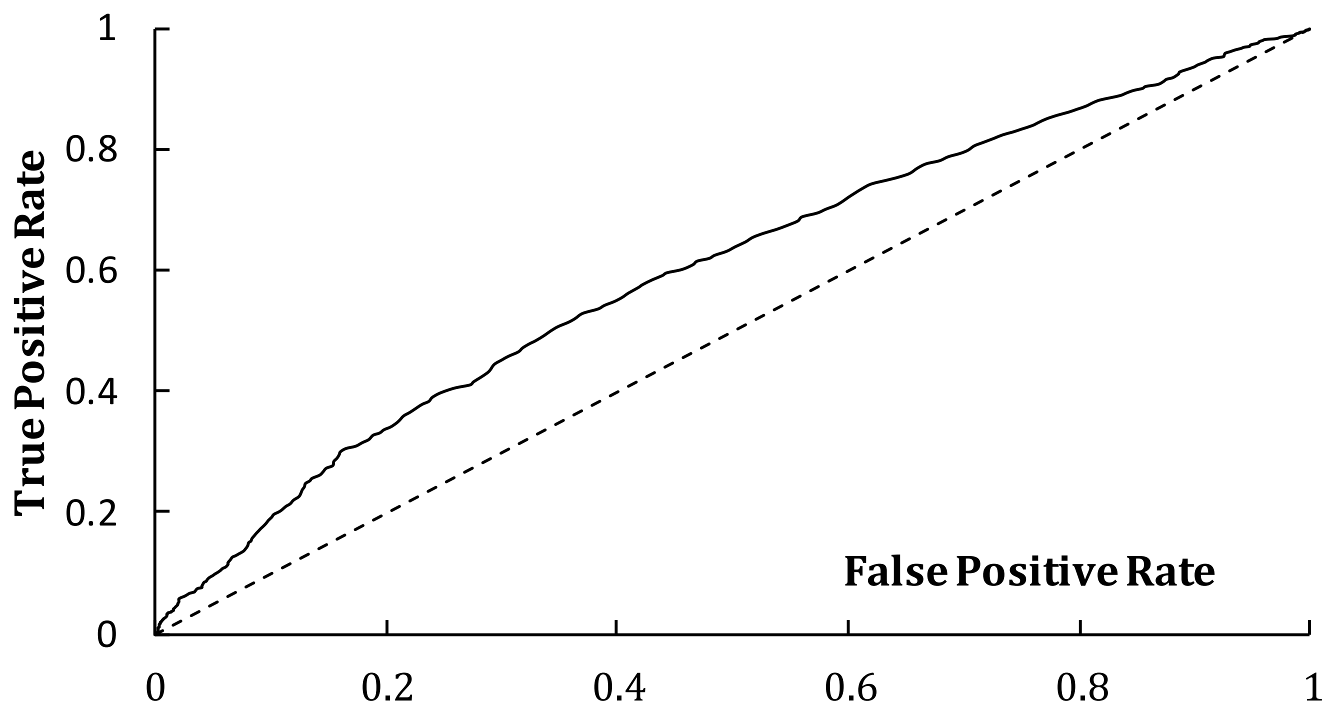
$$\text{Precision, P} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{F score} = 2 \times \frac{\text{P} \times \text{TPR}}{\text{P} + \text{TPR}}$$

Test Set Ratios for different Z values (Table 3.14)

	Z = 0.75	**Z = 0.80**	**Z = 0.85**
Accuracy	79.21%	63.47%	42.80%
True Positive Rate	94.48%	67.39%	34.89%
True Negative Rate	9.07%	45.46%	79.11%
False Positive Rate	90.93%	54.54%	20.89%
Precision	82.67%	85.02%	88.47%
F-score	88.18%	75.19%	50.04%

As we change the Z criterion we get an ROC curve (receiver operating characteristics) curve, Figure 3.11



Area Under Curve (AUC)

- The area under the curve is a popular way of summarizing the predictive ability of a model to estimate a binary variable
- When AUC =1 the model is perfect.
- When AUC =0.5 the model has no predictive ability
- When AUC<0.5 the model is worse than random
- In this case AUC = 0.6020

Choosing Z

- The value of Z can be based on
  - The expected profit from a loan that is good, P
  - The expected loss from a loan that defaults, L

- We need to maximize $P\times\mathrm{TP}-L\times\mathrm{FP}$

A Simple Alternative to regression : k-nearest neighbors

- Normalize data
- Measure the distance in *n*-dimensional space of the new data from the data for which there are labels (i.e. known outcomes)
- Distance of point with feature values x_i from point with feature values y_i is $\sqrt{\sum_i (x_i - y_i)^2}$
- Choose the *k* closest data items and average their labels
- For example if you are forecasting car sales in a certain area with *k*=3 and the three nearest neighbors for GDP growth and interest rates give sales of 5.2, 5.4 and 5.6 million units, the forecast would be the average of these or 5.4 million units.
- If you are forecasting whether a loan will default with *k*=5 and that of the five nearest neighbors four defaulted and one was good loan, you would estimate an 80% chance of default