150B/355B Introduction to Machine Learning for Social Science TA Section 5

Haemin Jee and Tongtong Zhang

February 8, 2018

Outline

Outline

1. Review of Regression
2. Review of Logistic Regression

Outline

1. Review of Regression
2. Review of Logistic Regression
3. Review of LASSO

Outline

1. Review of Regression
2. Review of Logistic Regression
3. Review of LASSO
4. Model Evaluation

Outline

Outline

Fundamentals

Q: What is a regression problem? What is a classification problem?

Fundamentals

Q: What is a regression problem? What is a classification problem?

A: In a regression, the dependent variable is (usually) a quantity. In a classification problem, the dependent variable is a label.

Fundamentals

Q: What is a regression problem? What is a classification problem?

A: In a regression, the dependent variable is (usually) a quantity. In a classification problem, the dependent variable is a label.

Q: What is inference? What is prediction?

Fundamentals

Q: What is a regression problem? What is a classification problem?

A: In a regression, the dependent variable is (usually) a quantity. In a classification problem, the dependent variable is a label.

Q: What is inference? What is prediction?

A: When doing inference, we want to say something about the relationship between some *independent variables* and the *dependent variable*. When doing prediction, we want to be able to predict the value of the dependent variable - do not really care about the relationship between the independent variables and the dependent variable.

Multivariate Regression

Basic Set Up:

Multivariate Regression

Basic Set Up:

$$Y_i = f(X_{1i}, X_{2i}, X_{3i}, X_{4i}) + \epsilon_i$$
$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{\beta}_4 X_{4i} + \epsilon_i$$

Multivariate Regression

Basic Set Up:

$$Y_i = f(X_{1i}, X_{2i}, X_{3i}, X_{4i}) + \epsilon_i$$
$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{\beta}_4 X_{4i} + \epsilon_i$$

Q: How do we interpret $\hat{\beta}_3$?

Multivariate Regression

Basic Set Up:

$$Y_i = f(X_{1i}, X_{2i}, X_{3i}, X_{4i}) + \epsilon_i$$
$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{\beta}_4 X_{4i} + \epsilon_i$$

Q: How do we interpret $\hat{\beta}_3$?

A: $\hat{\beta}_3$ is the change in Y_i associated with a unit increase in the value of X_1, holding all other X constant.

Predicting with MVR

Q: How do we then predict a new observation, using MVR?

Predicting with MVR

Q: How do we then predict a new observation, using MVR?

A: Use the values of the independent variables of the observation and the coefficients from the model.

Predicting with MVR

Q: How do we then predict a new observation, using MVR?

A: Use the values of the independent variables of the observation and the coefficients from the model.

Sometimes we use MVR when we have labels (or 0s and 1s as dependent variables). In those cases, we set a particular threshold and use that as a decision rule.

Logistic Regression Set Up

Goal: Predict a probability of an event or outcome happening, given some predictor variables.

Logistic Regression Set Up

Goal: Predict a probability of an event or outcome happening, given some predictor variables.

Let's call this p_i where i denotes an individual observation.

Logistic Regression Set Up

Goal: Predict a probability of an event or outcome happening, given some predictor variables.

Let's call this p_i where i denotes an individual observation.

$$\log(\frac{\hat{p}_i}{1-\hat{p}_i}) = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2$$

Logistic Regression Set Up

Goal: Predict a probability of an event or outcome happening, given some predictor variables.

Let's call this p_i where i denotes an individual observation.

$$\log(\frac{\hat{p}_i}{1-\hat{p}_i}) = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2$$

Instead of modeling the *probability* of an event as a linear function of predictor variables (as we did in the LPM), we are modeling the logit of p as a linear function of predictor variables.

Logistic Regression Set Up

Goal: Predict a probability of an event or outcome happening, given some predictor variables.

Let's call this p_i where i denotes an individual observation.

$$\log(\frac{\hat{p}_i}{1-\hat{p}_i}) = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2$$

Instead of modeling the *probability* of an event as a linear function of predictor variables (as we did in the LPM), we are modeling the logit of p as a linear function of predictor variables.

Q: How do we calcualte p?

Logistic Regression Set Up

Goal: Predict a probability of an event or outcome happening, given some predictor variables.

Let's call this p_i where i denotes an individual observation.

$$\log(\frac{\hat{p}_i}{1-\hat{p}_i}) = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2$$

Instead of modeling the *probability* of an event as a linear function of predictor variables (as we did in the LPM), we are modeling the logit of p as a linear function of predictor variables.

Q: How do we calcualte p?

A: We use the logistic function!

Interpreting Logistic Regression

Q: How do we interpret β_1 in a logistic regression?

Interpreting Logistic Regression

Q: How do we interpret β_1 in a logistic regression?

A: A one unit increase in the value of X_1 results in a β_1 increase / decrease in the *log odds* of an event happening.

LPM vs. Logistic Regression

Let's say you and co-authors are working on a research project on civil war. You want to know: what explains the onset of civil war? You collect a host of independent variables about countries: GDP, population, % of land that is mountainous terrain, ethnic fractionalization, etc. You also collect information about whether a civil war happened in that particular country in a particular year. Now, you're ready to run a model! But which one?

The LASSO

Q: What are some reasons we might want to use LASSO (as opposed to multivariate regression)?

The LASSO

Q: What are some reasons we might want to use LASSO (as opposed to multivariate regression)?

A:

- Too many predictors ($p > n$)

The LASSO

Q: What are some reasons we might want to use LASSO (as opposed to multivariate regression)?

A:

- Too many predictors ($p > n$)
- We have correlated predictors

The LASSO

Q: What are some reasons we might want to use LASSO (as opposed to multivariate regression)?

A:

- Too many predictors ($p > n$)
- We have correlated predictors
- Concerns of overfitting

LASSO Set Up

When using LASSO, we are using following cost function:

$$\beta_{\text{LASSO}} = \text{argmin}_{\beta} \sum_{i=1}^{N} (Y_i - \boldsymbol{\beta} \cdot \boldsymbol{x}_i)^2 + \lambda \sum_{p=1}^{P} |\beta_p|$$

Q: What is the λ parameter?

LASSO Set Up

When using LASSO, we are using following cost function:

$$\beta_{\text{LASSO}} = \text{argmin}_{\beta} \sum_{i=1}^{N} (Y_i - \boldsymbol{\beta} \cdot \boldsymbol{x}_i)^2 + \lambda \sum_{p=1}^{P} |\beta_p|$$

Q: What is the λ parameter?

A: The "tuning parameter" that tells us how much we want to penalize large coefficients.

Bias-Variance Trade Off

What kind of model do we want?

Bias-Variance Trade Off

What kind of model do we want?

- We want a model that captures correctly the systematic relationships in our data. (low bias)

Bias-Variance Trade Off

What kind of model do we want?

- We want a model that captures correctly the systematic relationships in our data. (low bias)
- We want a model that does not depend too much on our training data – ultimately we want to be able to apply this model to new observations. (low variance)

Bias-Variance Trade Off

What kind of model do we want?

- We want a model that captures correctly the systematic relationships in our data. (low bias)
- We want a model that does not depend too much on our training data – ultimately we want to be able to apply this model to new observations. (low variance)
- This is a common trade-off in many other machine learning algorithims

Bias-Variance Trade Off

What kind of model do we want?

- We want a model that captures correctly the systematic relationships in our data. (low bias)
- We want a model that does not depend too much on our training data – ultimately we want to be able to apply this model to new observations. (low variance)
- This is a common trade-off in many other machine learning algorithims

LASSO is *one* machine learning algorithim that is designed to protect against overfitting – and address this bias-variance trade off.

Bias-Variance Trade Off

What kind of model do we want?

- We want a model that captures correctly the systematic relationships in our data. (low bias)
- We want a model that does not depend too much on our training data – ultimately we want to be able to apply this model to new observations. (low variance)
- This is a common trade-off in many other machine learning algorithims

LASSO is *one* machine learning algorithim that is designed to protect against overfitting – and address this bias-variance trade off.

We introduce a bit of bias into the model (by penalizing the coefficients) so that we can decrease our variance and make **better predictions**.

Model Evaluation

Accuracy: All of our correct guesses divided by all guesses

Model Evaluation

Accuracy: All of our correct guesses divided by all guesses

$$\text{Accuracy} = \frac{\text{True Yes} + \text{True No}}{\text{True Yes} + \text{True No} + \text{False Yes} + \text{False No}}$$

Model Evaluation

Accuracy: All of our correct guesses divided by all guesses

$$\text{Accuracy} = \frac{\text{True Yes} + \text{True No}}{\text{True Yes} + \text{True No} + \text{False Yes} + \text{False No}}$$

Precision: Among all of our "Yes" guesses, how many were actually true "Yes"?

Model Evaluation

Accuracy: All of our correct guesses divided by all guesses

$$\text{Accuracy} = \frac{\text{True Yes} + \text{True No}}{\text{True Yes} + \text{True No} + \text{False Yes} + \text{False No}}$$

Precision: Among all of our "Yes" guesses, how many were actually true "Yes"?

$$\text{Precision} = \frac{\text{True Yes}}{\text{True Yes} + \text{False Yes}}$$

Model Evaluation

Accuracy: All of our correct guesses divided by all guesses

$$\text{Accuracy} = \frac{\text{True Yes} + \text{True No}}{\text{True Yes} + \text{True No} + \text{False Yes} + \text{False No}}$$

Precision: Among all of our "Yes" guesses, how many were actually true "Yes"?

$$\text{Precision} = \frac{\text{True Yes}}{\text{True Yes} + \text{False Yes}}$$

Recall: Among all the true Yes's, how many were we able to guess correctly?

Model Evaluation

Accuracy: All of our correct guesses divided by all guesses

$$\text{Accuracy} = \frac{\text{True Yes} + \text{True No}}{\text{True Yes} + \text{True No} + \text{False Yes} + \text{False No}}$$

Precision: Among all of our "Yes" guesses, how many were actually true "Yes"?

$$\text{Precision} = \frac{\text{True Yes}}{\text{True Yes} + \text{False Yes}}$$

Recall: Among all the true Yes's, how many were we able to guess correctly?

$$\text{Recall} = \frac{\text{True Yes}}{\text{True Yes} + \text{False No}}$$

Why Can't We (Just) Use Accuracy?

Let's assume we are in *binary classification* world, with a positive label and a negative label.

Why Can't We (Just) Use Accuracy?

Let's assume we are in *binary classification* world, with a positive label and a negative label.

When a classifier makes a prediction, there can be four possibilities:

Why Can't We (Just) Use Accuracy?

Let's assume we are in *binary classification* world, with a positive label and a negative label.

When a classifier makes a prediction, there can be four possibilities:

- True Positives (TP): the classifier labels an observation as positive and it really is positive

Why Can't We (Just) Use Accuracy?

Let's assume we are in *binary classification* world, with a positive label and a negative label.

When a classifier makes a prediction, there can be four possibilities:

- True Positives (TP): the classifier labels an observation as positive and it really is positive
- False Positives (FP): the classifier labels an observation as positive when it actually was negative

Why Can't We (Just) Use Accuracy?

Let's assume we are in *binary classification* world, with a positive label and a negative label.

When a classifier makes a prediction, there can be four possibilities:

- True Positives (TP): the classifier labels an observation as positive and it really is positive
- False Positives (FP): the classifier labels an observation as positive when it actually was negative
- True Negatives (TN): the classifier labels an observation as negative and it really is negative

Why Can't We (Just) Use Accuracy?

Let's assume we are in *binary classification* world, with a positive label and a negative label.

When a classifier makes a prediction, there can be four possibilities:

- True Positives (TP): the classifier labels an observation as positive and it really is positive
- False Positives (FP): the classifier labels an observation as positive when it actually was negative
- True Negatives (TN): the classifier labels an observation as negative and it really is negative
- False Negatives (FN): the classifier labels an observation as negative when it actually was positive

Why Can't We (Just) Use Accuracy?

Let's assume we are in *binary classification* world, with a positive label and a negative label.

When a classifier makes a prediction, there can be four possibilities:

- True Positives (TP): the classifier labels an observation as positive and it really is positive
- False Positives (FP): the classifier labels an observation as positive when it actually was negative
- True Negatives (TN): the classifier labels an observation as negative and it really is negative
- False Negatives (FN): the classifier labels an observation as negative when it actually was positive

The accuracy counts the total number of instances the classifier was *right* (true positives + true negatives) and divides it by the total number of classifications made.

An Example

Table: Example

	Real Positive	Real Negative
Classified Positive	**10 (TP)**	**25 (FP)**
Classified Negative	**15 (FN)**	**100 (TN)**

An Example

Table: Example

	Real Positive	Real Negative
Classified Positive	**10 (TP)**	**25 (FP)**
Classified Negative	**15 (FN)**	**100 (TN)**

$$\text{Accuracy} = \frac{10+100}{10+100+15+25} = .73$$

An Example

What happens when we switch to a non-informative classifier? Let's classify everything as a negative label.

Table: Example

	Real Positive	Real Negative
Classified Positive	**0 (TP)**	**0 (FP)**
Classified Negative	**25 (FN)**	**125 (TN)**

An Example

What happens when we switch to a non-informative classifier? Let's classify everything as a negative label.

Table: Example

	Real Positive	Real Negative
Classified Positive	**0 (TP)**	**0 (FP)**
Classified Negative	**25 (FN)**	**125 (TN)**

Accuracy $= \frac{0+125}{0+125+0+25} = .83$

An Example

What happens when we switch to a non-informative classifier? Let's classify everything as a negative label.

Table: Example

	Real Positive	Real Negative
Classified Positive	**0 (TP)**	**0 (FP)**
Classified Negative	**25 (FN)**	**125 (TN)**

Accuracy $= \frac{0+125}{0+125+0+25} = .83$

But this does not make sense! We used a completely non-informative model and still got an increase in accuracy!

An Example

This is called the **accuracy paradox**: if there are more false positives than true positives, we will always increase the accuracy if we change the classification rule to always predict a negative label. Similarly, when there are more false negatives than true negatives, we can increase our accuracy by always predicting a positive label.

Other Measures

Accuracy can be misleading; that's why we choose to use other metrics as well!

Other Measures

Accuracy can be misleading; that's why we choose to use other metrics as well!

Precision: Of the ones we classified as Positive, how many were True Positives?

Other Measures

Accuracy can be misleading; that's why we choose to use other metrics as well!

Precision: Of the ones we classified as Positive, how many were True Positives?

Recall: Of the True Positives that exist, how many were we able to classify correctly?

Other Measures

Accuracy can be misleading; that's why we choose to use other metrics as well!

Precision: Of the ones we classified as Positive, how many were True Positives?

Recall: Of the True Positives that exist, how many were we able to classify correctly?

Whether precision or recall matters more may depend on context

Other Measures

Accuracy can be misleading; that's why we choose to use other metrics as well!

Precision: Of the ones we classified as Positive, how many were True Positives?

Recall: Of the True Positives that exist, how many were we able to classify correctly?

Whether precision or recall matters more may depend on context

- If we are developing a system to detect fraud in bank transactions, we want to have a high recall (i.e. we want to correctly identify most of the fraud transactions that occur)
- If we are classifying pro and anti-Trump tweets, we may prefer to optimize precision (i.e. we want to make sure the tweets we've labeled as pro-Trump are actually so). False Negatives are not very consequential in this case and the source of data is so massive.