Data Science

Machine Learning Techniques

Lesson 6: Feature Engineering and Feature Selection

Hang Zhang, Ph.D.
November 5th, 2018

Outline of Lesson 6

- Capstone Project
  - Where we/you are?
  - Midterm report (2-pager report per team, due 11/11 midnight)
- Feature engineering
  - Categorical variables:
    - One-hot encoding
    - Risk values of categorical variables
  - Recency, Frequency, and Monetary (RFM) framework
- Feature selection
  - Filter based: Mutual Information
  - Step-wise: Forward, Backward, Both
  - Embedded: LASSO (L1 Regularization)

Capstone Project – Where Are You?

Team Name	Public Ranking	Best Score	# of Submissions
Datas R Us			
DS420_PandaPlayers	412	3.386	1
DS420_astroclass_capstone (Merged with datarangers)			
DS420_TYF			
DS420-GRJT	470	29.219	1
DS420_Galileo's_Gala			
DS420_Galaxy			

Capstone Project – Midterm Report

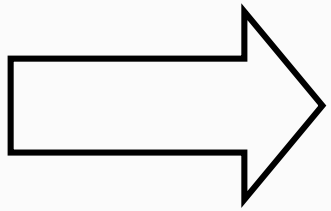
- 2-pager Word/PDF document from every team
  - Your team name
  - Your ranking on the leaderboard
  - What you have done:
    - What data you have been using
    - What features you have generated
    - What models you have built
  - What you plan to do next?
  - How do your team members collaborate?
- Due 11/11/2018 23:59 PST

Categorical Variables

- We call the number of unique values of a categorical variable the number of levels
- Categorical variables with high cardinality
  - A categorical variable has a large number of levels
- Challenges of categorical variables for scikit-learn machine learning models:
  - All variables have to be numerical
- Challenges of variables with high cardinality:
  - Random forest model in R can only handle at most 52 levels of a categorical variable

One hot Encoding

feature
red
blue
green
red
red
green
blue



red	blue	green
1	0	0
0	1	0
0	0	1
1	0	0
1	0	0
0	0	1
0	1	0

Dealing with Categorical Attributes

Outlook	Temperature	Humidity	Windy	Play
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	78	false	yes
rain	70	96	false	yes
rain	68	80	false	yes
rain	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rain	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rain	71	80	true	no

Attributes:

- Outlook (overcast, rain, sunny)
- Temperature real
- Humidity real
- Windy (true, false)
- Play (yes, no)

Standard Spreadsheet Format

OutLook	OutLook	OutLook	Temp	Humidity	Windy	Windy	Play	Play
overcast	rain	sunny			TRUE	FALSE	yes	no
0	0	1	85	85	0	1	1	0
0	0	1	80	90	1	0	0	1
1	0	0	83	78	0	1	1	0
0	1	0	70	96	0	1	1	0
0	1	0	68	80	0	1	1	0
0	1	0	65	70	1	0	0	1
1	0	0	64	65	1	0	1	0
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.

Problem of One Hot Encoding

- It significantly widens the dataset
  - If you have zip code as a feature in your dataset
  - There are approximately 43,000 zip codes in US
  - It means after one hot encoding, you will have 43,000 binary columns to represent zip codes
  - You may have other categorical variables…
  - Sometimes exceeds the memory limitation

Categorical Variable: Risk Values

- Calculate the risk value of each level of a categorical variable:

$$R_i = \log\left(\frac{\Pr(Y=1 \mid X=x_i)}{\Pr(Y=0 \mid X=x_i)}\right)$$

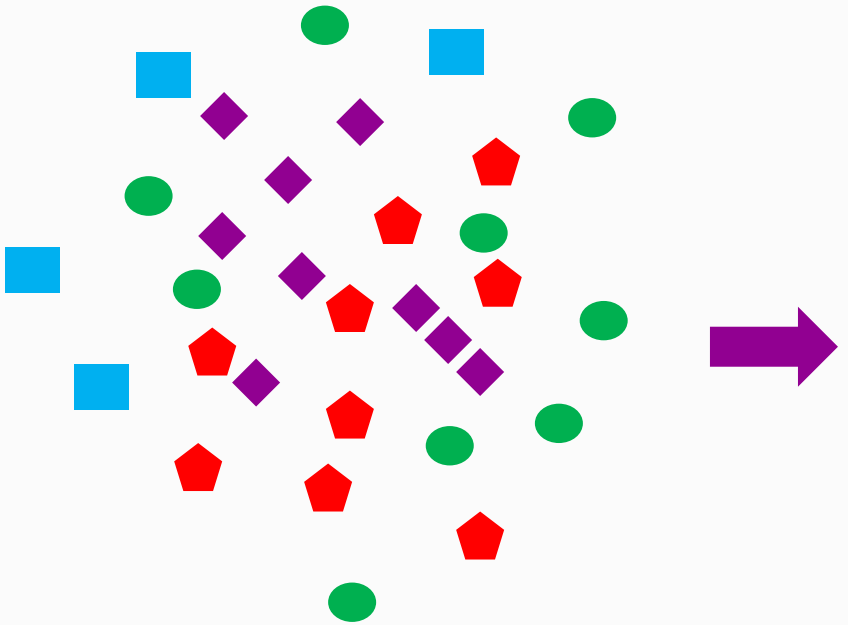
$$\Pr(Y=1 \mid X=x_i) \approx \frac{N_{Y=1 \& X=x_i}}{N_{X=x_i}}$$

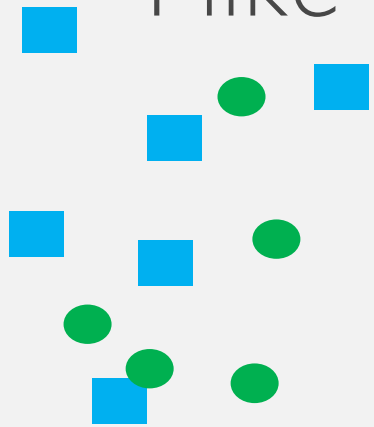
- Use risk values to replace categorical levels in the data
- Avoids widening the dataset
- Converts the categorical values to numerical values, preferable by many models
- The higher risk value of a level, the higher probability that the target variable = 1

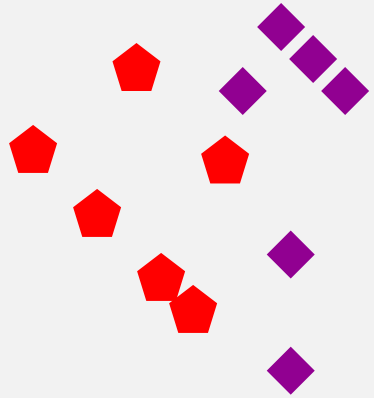
Feature Engineering

- Better to have a fair modeling method and good variables, than to have the best modeling method and poor variables
- It needs a lot of insights to the process that generates the data, and to the business it is going to be applied on
  - For instance, in fraud detection, you need to think about why sometimes you received calls from your credit card company about some early fraud alert?
  - In auto/home owner insurance, you need to think about why your insurance premium increased this year.
- It is the most time consuming, but also the most creative part of a data science project
  - The learned mapping between the target variable and the features can provide valuable insight to the business

Example of Feature Engineering







In-class Lab: Section 1

- Categorical Variables

Recency, Frequency, and Monetary

- A set of features very popularly used in customer churn problem, and other domains where analog exists
- Recency: How long has it passed since the most recent interaction between the user and the system
- Frequency: In the past given time period (1 month, 1 week, etc), how many times the user interacted with the system
- Monetary: In the past given time period, how much (time, money, etc) the user has spent on the system

Questions to Ask Before RFM Feature Engineering

- What is the definition of customer churn?
  - Option 1: When the activity of a customer in the next 1 week drops to < 10% of his/her regular usage
  - Option 2: When the activity of a customer in the next 1 week drops to < 30% of his/her regular usage
  - Which definition is better?

- When you want to determine which customers are churning (check point)?
  - Probably every day?

- Which group of customers you want to determine the probability of churn?
  - You always have new customers joining
  - You always have customers who do not have any activity for more than half a year
  - You can decide that every day, you want to detect the churn probability of customers who have activities in the past 6 months.
  - For customers who have no activities in the past 6 months, they already churned. No need to detect churn probability.
  - For customers who joined the web site after the check point, at the checkpoint, we do not know this customer yet, no need to detect churn probability.

Example of RFM Calculation

UserId	Age	Address	Column 0	Transactic	Timestamp	ItemId	Quantity	Value
1113	K	F	118152	904890	11/12/2000 0:00	4710000000000	2	29
1113	K	F	118153	905431	11/12/2000 0:00	4900000000000	3	391
1113	K	F	118154	1000113	11/26/2000 0:00	4900000000000	1	111
1113	K	F	118155	1000416	11/26/2000 0:00	7620000000000	1	268
1113	K	F	118156	1000417	11/26/2000 0:00	4710000000000	1	179
1113	K	F	118157	1018276	11/27/2000 0:00	4710000000000	1	14
1113	K	F	118158	1019142	11/27/2000 0:00	4720000000000	1	224
1113	K	F	118159	1019267	11/27/2000 0:00	4710000000000	1	65
1113	K	F	118160	1019384	11/27/2000 0:00	4710000000000	1	116
1113	K	F	118161	1019478	11/27/2000 0:00	4710000000000	1	116
1113	K	F	118162	1019482	11/27/2000 0:00	4710000000000	1	89
1113	K	F	118163	1282039	1/6/2001 0:00	4710000000000	1	188
1113	K	F	118164	1284131	1/6/2001 0:00	4710000000000	1	28
1113	K	F	118165	1284189	1/6/2001 0:00	4710000000000	2	84
1113	K	F	118166	1284585	1/6/2001 0:00	37000440147	1	47
1113	K	F	118167	1284765	1/6/2001 0:00	4900000000000	1	169
1113	K	F	118168	1284951	1/6/2001 0:00	9560000000000	1	28
1113	K	F	118169	1285516	1/6/2001 0:00	4710000000000	2	84
1250	D	D	243280	1494035	2/4/2001 0:00	4720000000000	1	148
1250	D	D	243281	1494721	2/4/2001 0:00	4720000000000	1	179
1250	D	D	243282	1494852	2/4/2001 0:00	4910000000000	1	309
1250	D	D	243283	1495078	2/4/2001 0:00	4720000000000	2	98
1250	D	D	243270	1451064	2/10/2001 0:00	4710000000000	1	89
1250	D	D	243271	1451293	2/10/2001 0:00	4710000000000	1	65
1250	D	D	243272	1451301	2/10/2001 0:00	723000000000	1	65
1250	D	D	243273	1451534	2/10/2001 0:00	20480349	1	395
1250	D	D	243274	1451641	2/10/2001 0:00	4710000000000	2	44
1250	D	D	243275	1451863	2/10/2001 0:00	4710000000000	1	28
1250	D	D	243276	1452120	2/10/2001 0:00	4710000000000	2	26
1250	D	D	243277	1452219	2/10/2001 0:00	4710000000000	2	44
1250	D	D	243278	1452444	2/10/2001 0:00	4710000000000	1	28
1250	D	D	243279	1452672	2/10/2001 0:00	4710000000000	1	65

- If we set the checkpoint 11/20/2000
- Recency: 11/12 – 11/20 = 8 days
- Frequency and monetary in the past 2 weeks
  - Frequency: 2 (2 transactions)
  - Monetary Value: 29+391 = 420
  - Monetary Quantity: 2+3 = 5

- If we want the frequency and monetary in the most recent 7 days:
- Frequency = 0
- Monetary Value and Quantity = 0

In-class Lab: Section 2

- RFM

Feature Selection

- Process of selecting a subset of features that are good predictors of the target

- Useful for
  - Controlling complexity of model
  - Speeding up model learning without reducing accuracy
  - Improving generalization capability

Model Selection vs Feature Selection

- Model selection includes selecting:
  - Model algorithm
  - Model algorithm hyperparameters
  - Features to be used to train the models

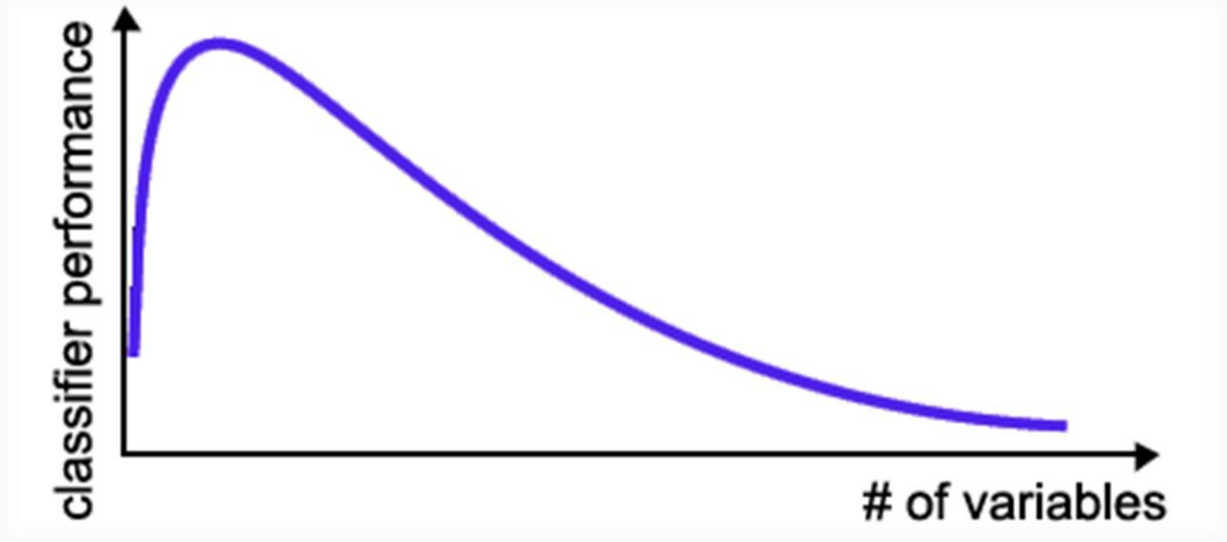
- Feature selection
  - Select features to be used to train the models

Why We Need Feature Selection?

Curse of Dimensionality

- The required number of samples (to achieve the same accuracy) grows exponentially with the number of variables!
- In practice: number of training examples is fixed!

  the classifier's performance will degrade for a large number of features!



In many cases the information lost by discarding variables is made up for by a more accurate mapping/sampling in the lower-dimensional space !

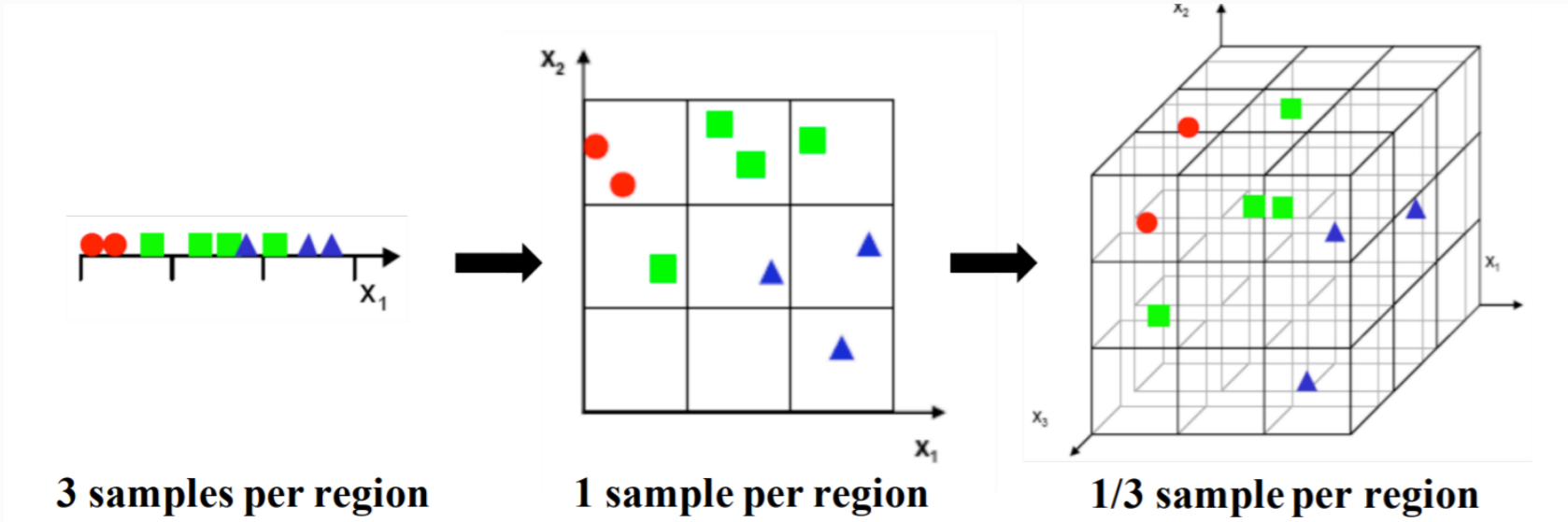
Problems of High-Dimensional Data

- High-dimensional data is often notorious to tackle due to the curse of dimensionality
  - Increase storage and running time
  - Overfit the machine learning models
  - Require more data

- The intrinsic dimension of data may be small
  - The number of genes responsible for a certain disease

Curse of Dimensionality – Required Samples

- Data sparsity becomes exponentially worse as feature dimension increases
  - Conventional distance metrics become ineffective
  - All points in the high-dimension space look equally distant



http://nikhilbuduma.com/2015/03/10/the-curse-of-dimensionality/

Feature Selection, 3 types of methods

Filter Methods, select a subset of features before training a model, e.g.

- Correlation with target,
- Mutual Information between feature and target

- *Simple to implement, and have reasonable performance*

Wrapper Methods, search combination of feature space by training and evaluating model using a subset of features, e.g.

- Forward, backward, step-wise feature selection,
- Genetic algorithms.

- *Computationally expensive and prone to over-fitting*

Embedded Methods, feature subset is chosen as part of model training, e.g.

- LASSO (L-1) regression, Regularized decision trees, random forests

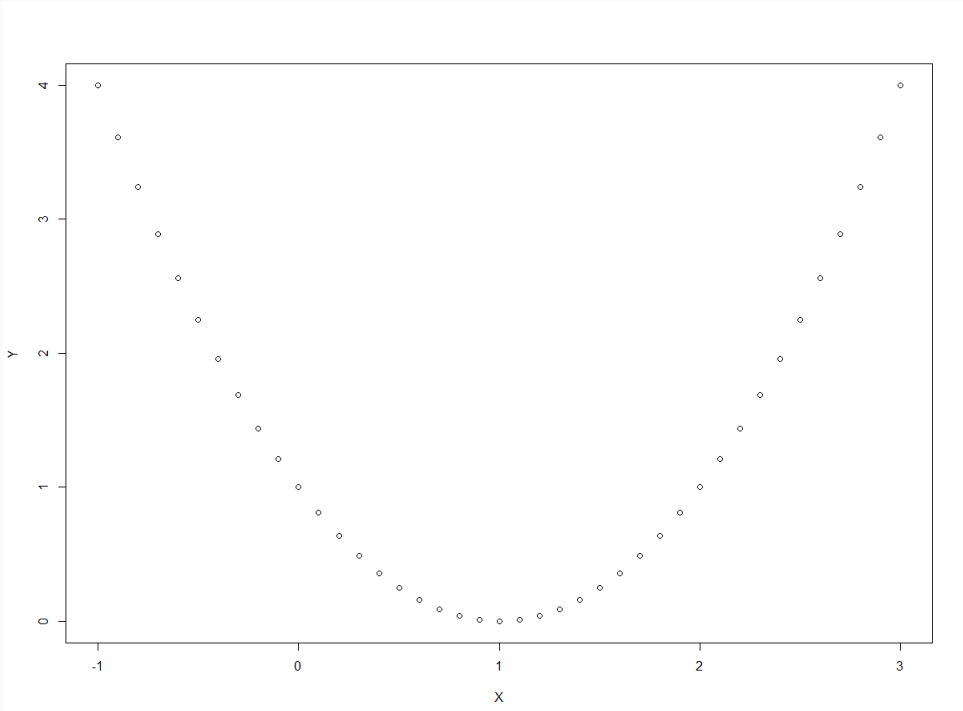
- *Typically robust to over-fitting, but has hyper parameters that will need to be fit using a validation data*

Filter-based Feature Selection

- Correlation with target variable
  - A good starting point
  - If Y is categorical variable (classification):
    - Use chi-square test to decide the correlation between each categorical X variable and Y variable
    - Use ANOVA test to decide the correlation between each numerical X variable and Y variable
  - If Y is continuous variable (regression):
    - Use ANOVA test to decide the correlation between each categorical X variable and Y variable
    - Use correlation between each numerical X variable and Y variable
  - *Alert*: If x1 and x2 are highly correlated, and x1 and Y are highly correlated, both x1 and x2 will be selected based on correlation with Y. Strong correlations in X will bring some challenge for some machine learning models, such as linear regression model.

Is Correlation Always a Good Choice?

- It makes sense for linear regression (logistic regression) model.
  - Since linear regression model only looks at linear relationship
- Does not make sense for nonlinear models such as tree-based models
- Cannot capture nonlinear relationship between X and Y



Mutual Information

- Captures Statistical Dependency between Two Variables
  - If two variables are statistically independent

$$\Pr(X, Y) = \Pr(X) \times \Pr(Y)$$

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)\,p(y)} \right)$$

  - Estimate Pr(X) from observations by using a kernel function

$$\hat{f}(x) = \frac{1}{Nh\sqrt{2\pi}} \sum_{i=1}^{N} \exp\left(\frac{-(x - x_i)^2}{2h^2}\right).$$

Step-wise Model (Feature) Selection

- Forward:
  - Start with a model with only inception
  - Add one feature in the model at each step
  - At each step, the variable that can maximally reduce the residual sum of squares (RSS) is chosen as the feature to add in the model.
- Backward:
  - Start with a model with all features
  - Remove one feature from the model at each step
  - At each step, the variable that can minimally increase the residual sum of squares (RSS) is chosen as the feature to remove from the model.
- Both:
  - At each step, will check whether add a feature, or remove a feature

How to Select the Best Model (Feature Set)?

- Akaike information criterion (AIC)
  - k: number of coefficients to estimate in the model
  - L: likelihood of the training data based on the model

$$\mathrm{AIC} = 2k - 2\ln(\hat{L})$$

- Bayesian information criterion

$$\mathrm{BIC} = \ln(n)k - 2\ln(\hat{L}).$$

- Choose the model that has the minimal AIC or BIC
- AIC tends to choose a larger model than BIC
  - AIC has less penalty on the complexity of model (k) than BIC

Embedded Method

- Lasso (least absolute shrinkage and selection operator)

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{N} \|y - X\beta\|_2^2 \right\} \text{ subject to } \|\beta\|_1 \leq t.$$

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{N} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$$

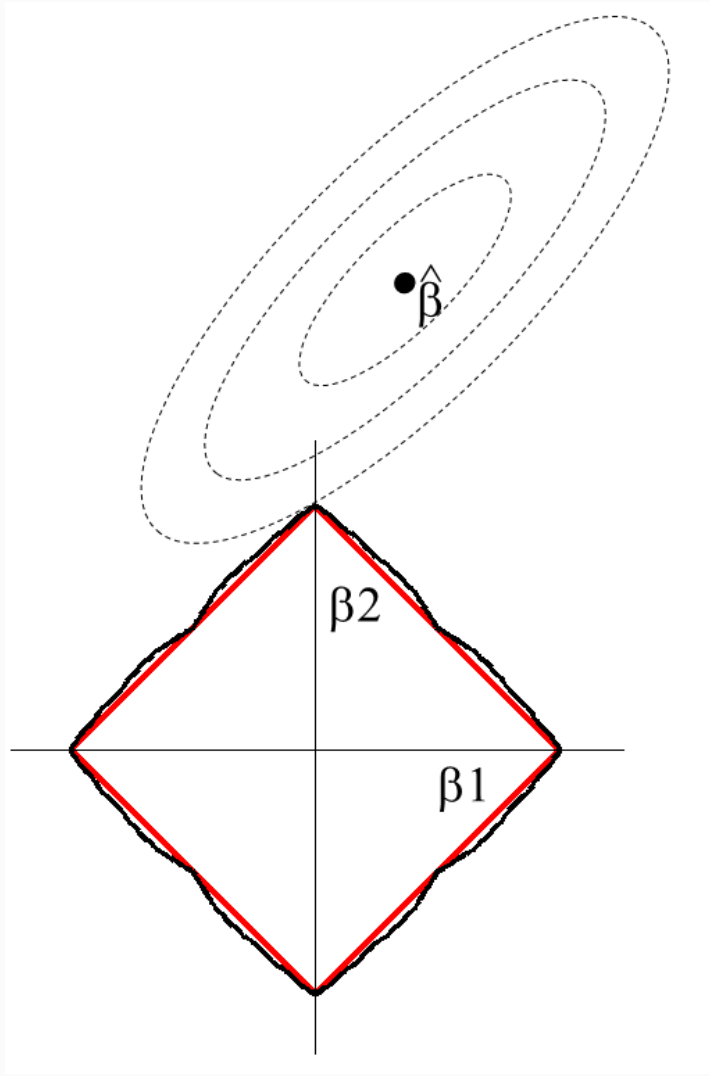
- Based on the second equation, we are penalizing on the complexity of the model (The sum of absolute values of the coefficients)

Why LASSO Can Select Features?

- Assuming only 2 X variables
- $\hat{\boldsymbol{\beta}}$ is the coefficient vector where there is no penalty
- Ellipsoid is the contour of MSE when coefficients change
- Very likely, some contour will meet with $|\beta_1| + |\beta_2| \leq t$

At the corner

- At the corner, the coefficients of some variables are set to 0
- These variables are de-selected



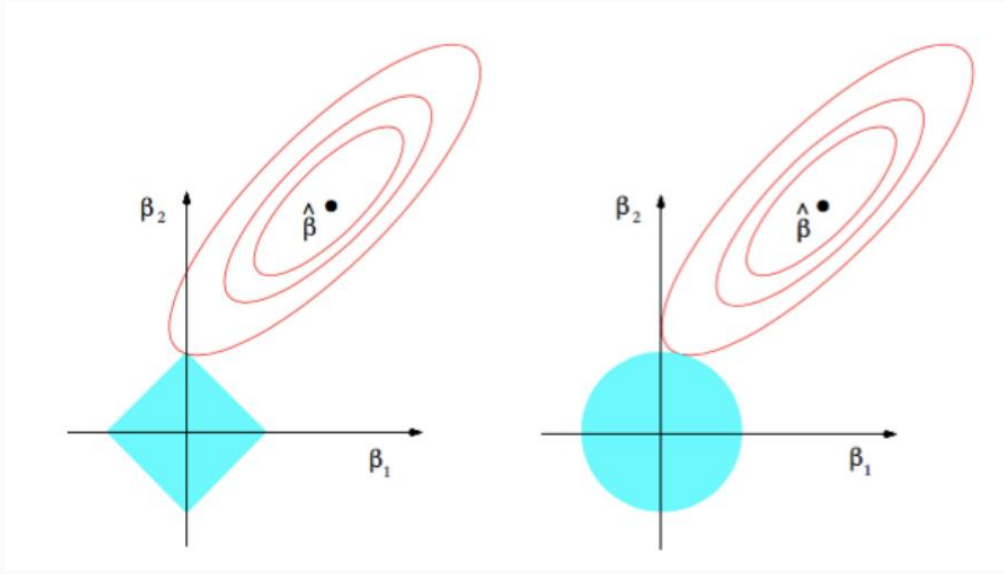
LASSO and Ridge Regression

- Ridge Regression

$$\text{minimize} \sum_{i=1}^{n}(y_i - \beta^\top \mathbf{z}_i)^2 \text{ s.t. } \sum_{j=1}^{p}\beta_j^2 \leq t$$

- Ridge Regression can be helpful when Z is highly correlated
  - $(Z^TZ)^{-1}$ does not exist, or is very sensitive to noise
  - $(Z^TZ + \lambda I_p)$ is always inversible.
- But Ridge Regression just shrinks variables, it does not select variables

$$\hat{\beta}_\lambda^{\text{ridge}} = (\mathbf{Z}^\top\mathbf{Z} + \lambda\mathbf{I}_p)^{-1}\mathbf{Z}^\top\mathbf{y}$$



Feature Selection and Engineering

Optimality?

In theory the goal is to find an optimal set of features, one that maximizes the scoring function…

In real world applications this is usually not possible

- For most problems it is computationally intractable to search the whole space of possible feature subsets
- One usually has to settle for approximations of the optimal subset
- Most of the research in this area is devoted to finding efficient search-heuristics

In-class Lab: Section 3

- Feature Selection