NYU CDS, SPRING 2018

INTRODUCTION TO DATA SCIENCE

FEATURE SELECTION

- Wide datasets: # of features p > # of examples n
  - Genomics: measure expression of many genes
  - Natural language models: word representation
- Identify good subset of features

BEST SUBSET REGRESSION

1. Start with $m = 0$ and the null model $\hat{\eta}_0(x) = \hat{\beta}_0$, estimated by the mean of the y_i.
2. At step $m = 1$, pick the single variable j that fits the response best, in terms of the loss L evaluated on the training data, in a univariate regression $\hat{\eta}_1(x) = \hat{\beta}_0 + x_j' \hat{\beta}_j$. Set $\mathcal{A}_1 = \{j\}$.
3. For each subset size $m \in \{2, 3, \ldots, M\}$ (with $M \leq \min(n-1, p)$) identify the best subset $\mathcal{A}_m$ of size m when fitting a linear model $\hat{\eta}_m(x) = \hat{\beta}_0 + x'_{\mathcal{A}_m} \hat{\beta}_{\mathcal{A}_m}$ with m of the p variables, in terms of the loss L.
4. Use some external data or other means to select the "best" amongst these M models.

BEST SUBSET REGRESSION

- All subsets
  - For p features, 2^p combinations
  - NP complete

BEST SUBSET REGRESSION

$$\min \|x\|_0 \text{ subject to } y = Ax$$

FORWARD STEPWISE REGRESSION

1. Start with $m = 0$ and the null model $\hat{\eta}_0(x) = \hat{\beta}_0$, estimated by the mean of the y_i.
2. At step $m = 1$, pick the single variable j that fits the response best, in terms of the loss L evaluated on the training data, in a univariate regression $\hat{\eta}_1(x) = \hat{\beta}_0 + x_j' \hat{\beta}_j$. Set $\mathcal{A}_1 = \{j\}$.
3. For each subset size $m \in \{2, 3, \ldots, M\}$ (with $M \leq \min(n-1, p)$) identify the variable k that when augmented with $\mathcal{A}_{m-1}$ to form $\mathcal{A}_m$, leads to the model $\hat{\eta}_m(x) = \hat{\beta}_0 + x'_{\mathcal{A}_m} \hat{\beta}_{\mathcal{A}_m}$ that performs best in terms of the loss L.
4. Use some external data or other means to select the "best" amongst these M models.

FORWARD STEPWISE REGRESSION

- Greedy
- May not be optimal

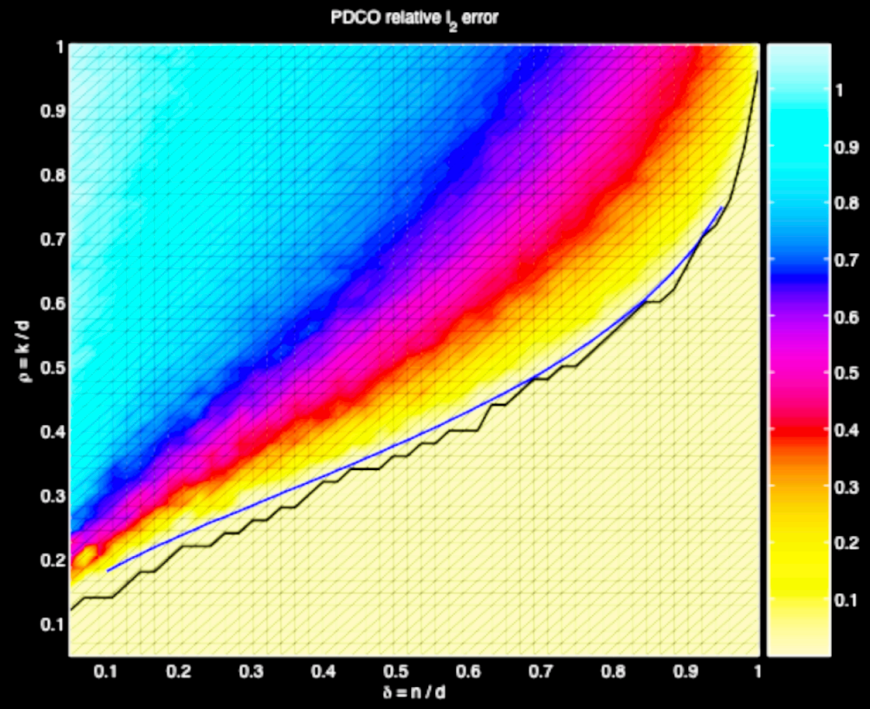
SPARSITY

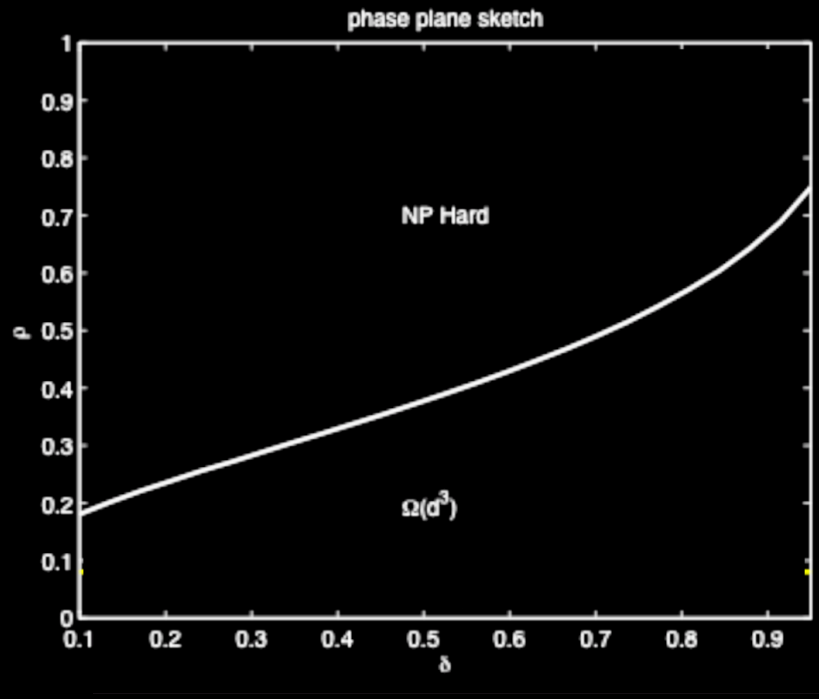
$$\min \|x\|_0 \text{ subject to } y = Ax$$

$$\min \|x\|_1 \text{ subject to } y = Ax$$

SPARSITY PHASE TRANSITION

- Y-axis: # non-zero coefficients / # examples
- X-axis: # of features / # of examples





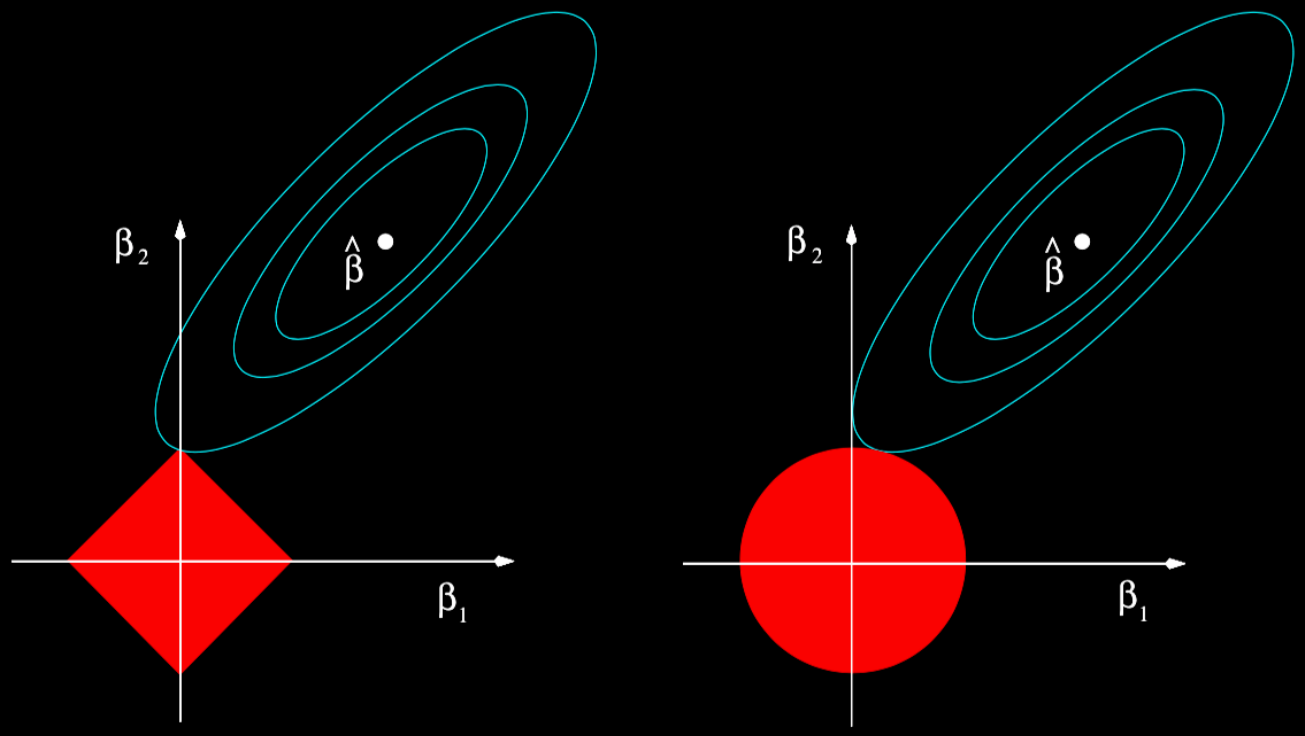
$$\|x - x_0\|_2 / \|x_0\|_2$$

LASSO

$$\min \|y - Ax\|_2^2 \text{ subject to } \|x\|_1 \leq q.$$

$$\min \|y - Ax\|_2^2/2 + \lambda\|x\|_1$$

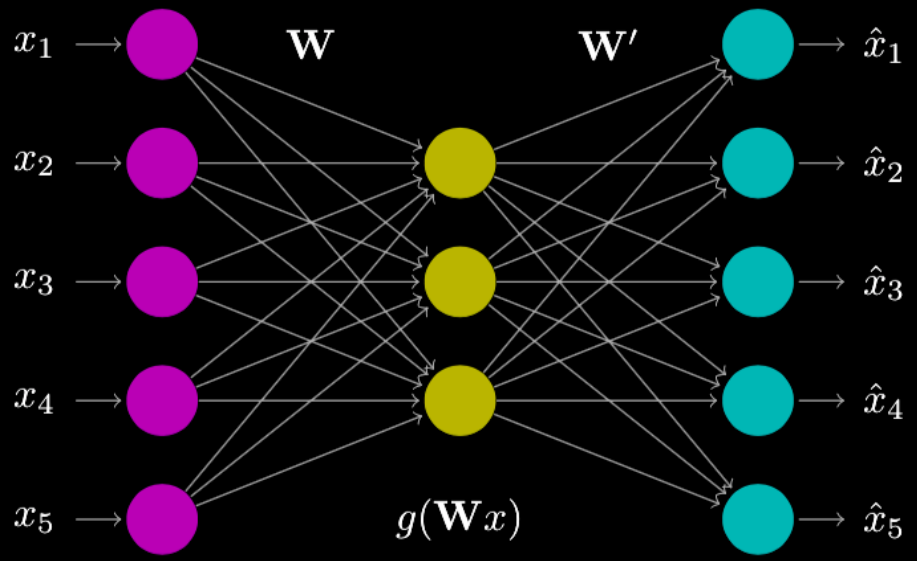
SPARSITY



NON-LINEAR PCA

$$\underset{\boldsymbol{A}\in\mathbb{R}^{p\times q},\ \boldsymbol{A}'\boldsymbol{A}=\boldsymbol{I}_q}{\text{minimize}} \sum_{i=1}^{n} \|x_i - \boldsymbol{A}\boldsymbol{A}'x_i\|_2^2$$

$$\underset{\boldsymbol{W}\in\mathbb{R}^{q\times p}}{\text{minimize}} \sum_{i=1}^{n} \|x_i - \boldsymbol{W}'g(\boldsymbol{W}x_i)\|_2^2$$



FEATURE EXTRACTION

- Training data x
- Learning finds w by optimization
- Feature extraction finds phi(x) by prior domain knowledge
- Feature extractor phi should select features useful for prediction
- Feature extractor defines a hypothesis class, model family, of possible predictors
- Learning is choosing a predictor from that space based on data

FEATURE REPRESENTATION

- Array: suitable for dense features
- Map: suitable for sparse features

REFERENCES

- Reading: CASI Ch 16

MIDTERM

- In class March 7th.
- 4-5 questions
- Read DSB and accompanying jupyter workbooks

PRACTICE QUESTION 1

- You are constructing an information-gain-based decision tree to predict a dependent variable “loan payment,” based on bank records. “Loan payment” has two possible values: “yes” and “no”
- Your dataset also includes the following variables: “Balance”, “Residence” (privately owned or rent), “age”, “gender”, “employment”, “marital status”
- After beginning to build the tree, you reached the following tree structure:

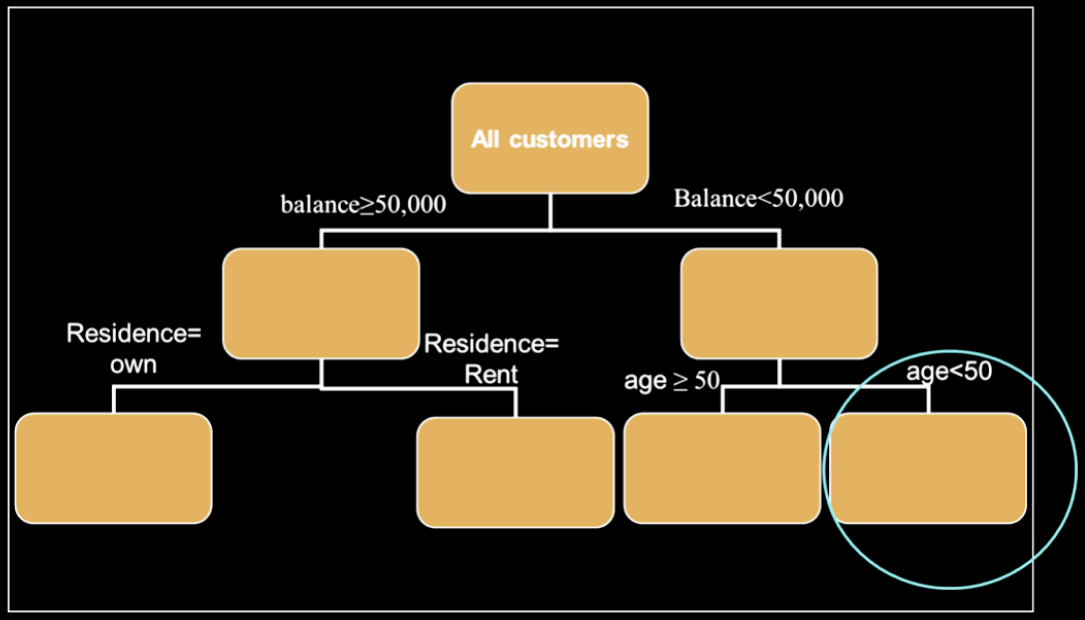
- Which variables would you consider as candidates to be used for additional splitting of the node marked by the circle?

  a. “Balance”, “Residence”, “Age”, “Gender”

  b. “Residence”, “Gender”, “Employment”, “Marital status”

  c. “Gender”, “Employment”, “Marital status”

  d. All, “Balance”, “Residence”, “Age”, “Gender”, “Employment”, “Marital status”



PRACTICE QUESTION 2

- A healthcare provider developed a model to identify people at high risk for developing a severe heart condition, based on their medical history. The model classifies individuals into two categories:

1. "positive" – expected to develop a severe heart condition
2. "negative" – not expected to develop a severe heart condition

- Based on this model, the healthcare provider is planning to invite people who were classified "positive" to undergo additional medical examinations. If the additional medical examinations confirm a high likelihood to develop a severe heart condition, preventive treatment could begin immediately. The treatment has a very good chance of success. The cost of the additional medical examination is relatively low. Given that the healthcare provider wishes to prevent as many people as possible from developing severe heart conditions, what will the healthcare provider prefer to obtain from this model:

a. High precision for the "positive" class

b. High recall for the "positive" class

c. The healthcare provider is indifferent to high precision or high recall for the positive "class"

SENSITIVITY

- Sensitivity refers to test's ability to correctly detect ill patients who have the condition.
- In a medical test used to identify a disease, sensitivity of test is proportion of people who test positive for disease among those who have the disease.
- TP / (TP + FN) = TP / total # of sick people in population

PRACTICE QUESTION 3

- A university has extensive dataset on its alumni, including past studies, demographic information by zip code, and past donations. The university is planning to send a deluxe brochure and a donation request to some of the alumni (Total targeting cost is $15 per individual) and has sufficient budget for constructing targeting models and running experiments under the following assumptions:
- Donation amount may vary.
- Alumni may spontaneously make a donation (even when not targeted).
- Targeting cost is fixed (C= $15).
- Other than the targeting cost, there are no additional costs for alumni who are targeted and decide not to donate.
- You have been asked to build several data mining models that would suggest which alumni should be targeted. Use the expected value framework to determine which models should be used to address the problem.
- Note: It is sufficient to write down the correct expected value equations to identify the models that should be constructed. There is no need to further solve/develop the equations.

EXPECTED VALUE

- Expected benefit of targeting:
- P(Donate|x,Targ)*[U(Donate|x,reponse)-C]+ [1-P(Donate|x,Targ)]*(-C)
- Expected benefit of NOT targeting:
- P(Donate|x, NotTarg)* U(Donate|x, NotTarg)
- Therefore, the following models should be constructed:

A classification model for P(Donate|x,Targ) – the probability that an individual, with a feature vector x, will donate, given he/she was targeted.

A regression model for U(Donate|x,response) – the amount of money donated by an individual with a feature vector x, given he/she responds to a targeted donation request.

A classification model for P(Donate|x, NotTarg) – the probability that an individual with a feature vector x, will donate given he/she was not targeted (spontaneous donation).

A regression model for U(Donate|x, NotTarg) – the amount of money donated by an alumnus, with a feature vector x, who makes a spontaneous donation.

PRACTICE QUESTION 4

Y\X	1	2	3	4	
1	1/8	1/16	1/32	1/32	1/4
2	1/16	1/8	1/32	1/32	1/4
3	1/16	1/16	1/16	1/16	1/4
4	1/4	0	0	0	1/4
	1/2	1/4	1/8	1/8	

MUTUAL INFORMATION OF X;Y

- H(X) = - sum_x p(x) log p(x) = -1/2log1/2-1/4log1/4-1/8log1/8-18/log1/8
- H(Y) = -sum_y p(y) log p(y) = -4 1/4 log1/4
- H(X|Y) = sum_i p(Y=i) H(X|Y=i) = 1/4 H(1/2,1/4,1/8,1/8) + 1/4 H(1/4, 1/2, 1/8, 1/8) + 1/4 H(1/4,1/4, 1/4,1/4) + 1/4H(1,0,0,0)
- I(X;Y) = H(X) - H(X|Y)
- Reduction in uncertainty of X due to knowledge of Y

DSB JUPYTER NOTEBOOKS

- https://github.com/mariazm/Spring2017_ProfFosterProvost