Final exam review

C S 487/519

Venue & Time

- Tuesday May 3, 2022
- 1:00pm - 3:00pm
- M02: Come to class
- M70: Take it online (Lockdown browser + camera)

How to take the exam

- Calculator is allowed
- Cell phone is NOT allowed
- 1 page cheat sheet (letter-size), allowing text on one-side or both-sided, hand-written or printed
- Plenty of blank paper
- Pencil/pen
- 2 hrs (online with 15 minutes extra to accommodate technical issues)

Question types

- Multi-choice questions
- True/False questions
- Short answer questions
- Calculation questions
- Programming questions (NO)

Scope (1)

- Introduction
- Linear Neural Networks
  - Perceptron
  - Adaline
  - SGD
- Classification
  - Support Vector Machine
  - Decision trees
  - Logistic Regression
  - Model evaluation
  - Model diagnose and parameter tuning

Scope (2)

- Dimension Reduction
  - PCA
  - LDA
  - Kernel PCA
- Regression
  - Linear
  - Non-linear
- Clustering
  - K-means
  - Hierarchical

Scope (3)

- Ensemble
  - Bagging
  - Random Forest
  - AdaBoost
- CNN
  - Only theory
- RNN
  - Only theory

Introduction – Exercise 1

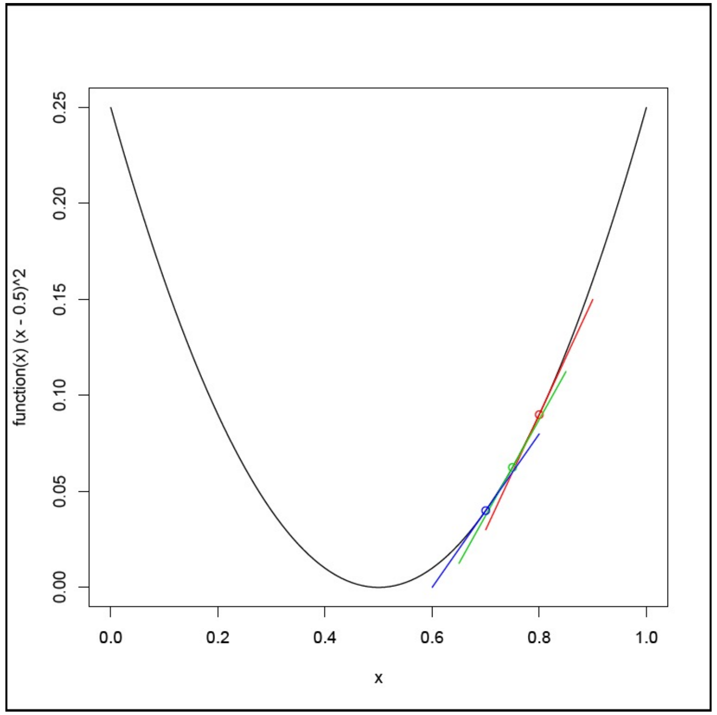
- Which of the following is not a machine learning task?
  a) Clustering
  b) Reinforcement learning
  c) Classification
  d) Principal Component Analysis
  e) Searching for a phone number from a phone book

Introduction - Exercise 2

- What are the major differences between the classification problem and the regression problem?
  - For the classification problem, the predicted value is categorical. For regression problem, the predicted value is numerical.

Linear NN – Exercises 1&2

- Given a function in the figure, which point has the largest gradient?
  - (a) red
  - (b) green
  - (c) blue
- What are the following steps not necessary when doing SGD?
  - (a) Data shuffling
  - (b) Using adaptive learning rate
  - (c) Updating weights



Linear NN – Exercise 3

- Give one major difference between the perceptron and the Adaline model.
  - In Perceptron, weights are updated using a unit step function.

$$\varphi(z) = \begin{cases} 1\ if\ z \geq 0 \\ -1\ otherwise \end{cases}$$

  - In Adaline, weights are updated based on a linear activation function

$$\varphi(z) = \phi(\boldsymbol{w}^{\top}\boldsymbol{x}) = \boldsymbol{w}^{\top}\boldsymbol{x}$$

Linear NN – Exercise 4 (update weights of linear NN models)

- Given a dataset with 100 instances and each instance has two features and one class label. Assume that we learn a perceptron model. Answer the following questions
- (1) How many values in the weight vector?
- (2) Given an instance **x** = (2, 3) with class label 1, learning rate η=0.01, and the value of each initial weight parameter be 0.1.
  - (a) Assume that one iteration of the perceptron predicts the label of x be 1, what will the Δw look like?
  - (a) Assume that one iteration of the perceptron predicts the label of x be -1, what will the Δw look like?

Solution ideas

- (1) 3 weight values, w = (w_0, w_1, w_2)
- (2.a) The initial w = (0.1, 0.1, 0.1). The updating equations are
  - $\Delta w_0 = \eta\left(y^{(i)} - \hat{y}^{(i)}\right)$ because $x_0^{(i)}$=1
  - $\Delta w_1 = \eta\left(y^{(i)} - \hat{y}^{(i)}\right)x_1^{(i)}$
  - $\Delta w_2 = \eta\left(y^{(i)} - \hat{y}^{(i)}\right)x_2^{(i)}$

  The prediction is correct. Thus, $y^{(i)} - \hat{y}^{(i)}$=0, $\Delta w_0 = \Delta w_1$= Δw_2 = 0
- (2.b) The prediction is wrong, $y^{(i)} - \hat{y}^{(i)}$=1-(-1)=2
  - $\Delta w_0 = \eta\left(y^{(i)} - \hat{y}^{(i)}\right)$ = 0.01*2=0.02
  - $\Delta w_1 = \eta\left(y^{(i)} - \hat{y}^{(i)}\right)x_1^{(i)}$= 0.01*2 * 2 = 0.04
  - $\Delta w_2 = \eta\left(y^{(i)} - \hat{y}^{(i)}\right)x_2^{(i)}$ = 0.01*2 * 3 = 0.06

Classification - Exercise 1

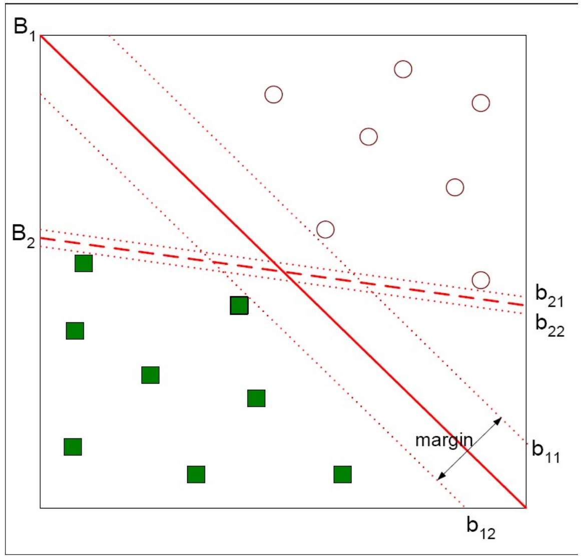
- Why do we need feature scaling?

- Give one feature scaling method.

$$x'_j = \frac{x_j - \mu_j}{\sigma_j}$$

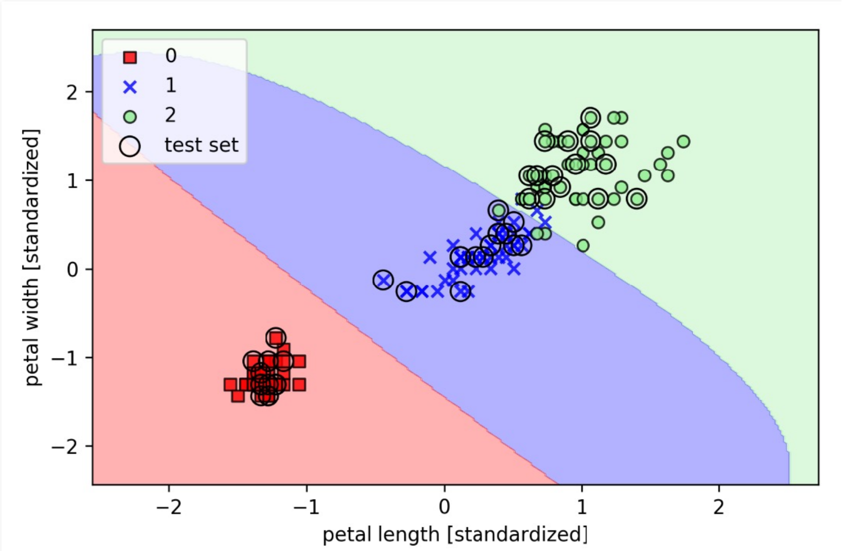
Classification - Exercise 2

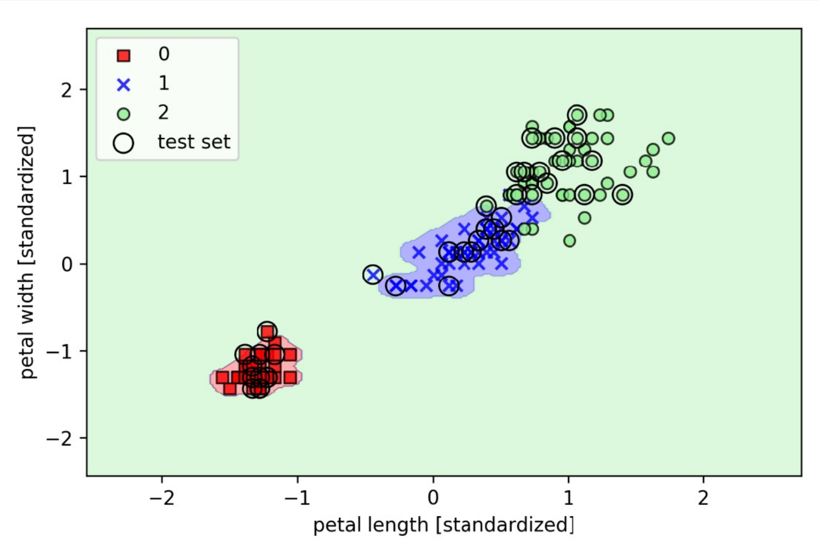
- Assume that I learn two linear SVM models B1 and B2 (see figure below) to separate the green and red points. Which model will you choose and why?



Classification – Exercise 3

- If we learn two kernel SVM models to classify the red, blue, and green points below. Which model tends to overfit and why?





Classification - Exercise 4 (SVM)

- For a SVM, assume that we learned w11=-10, w2 = -10, and b=10. Give a new instance (0.1, 0.3), what’s your prediction?

Solution steps

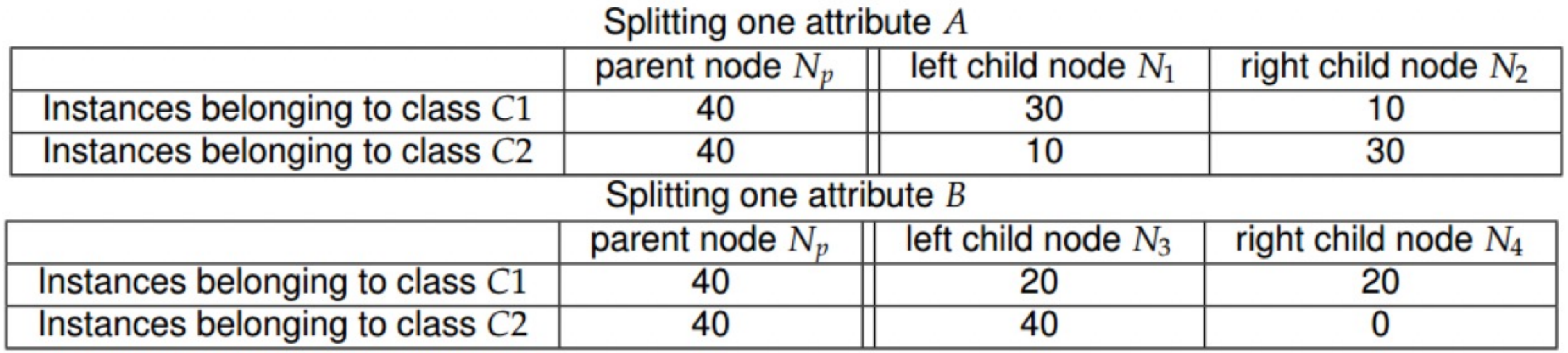
$w^T z$ +b = (-10,-10) $\begin{pmatrix} 0.1 \\ 0.3 \end{pmatrix}$ +10 = -4 + 10 = 6

Sign is positive

Predict this instance to be positive.

Classification – Exercise 5

- (Decision trees) Given the following data statistics of a decision tree node with 80 instances where 40 belong to class C1 and 40 instances belong to class C2. Each instance has two attributes A and B. If you use Gini index to calculate node impurity. Which attribute will you use to do the splitting? And why.



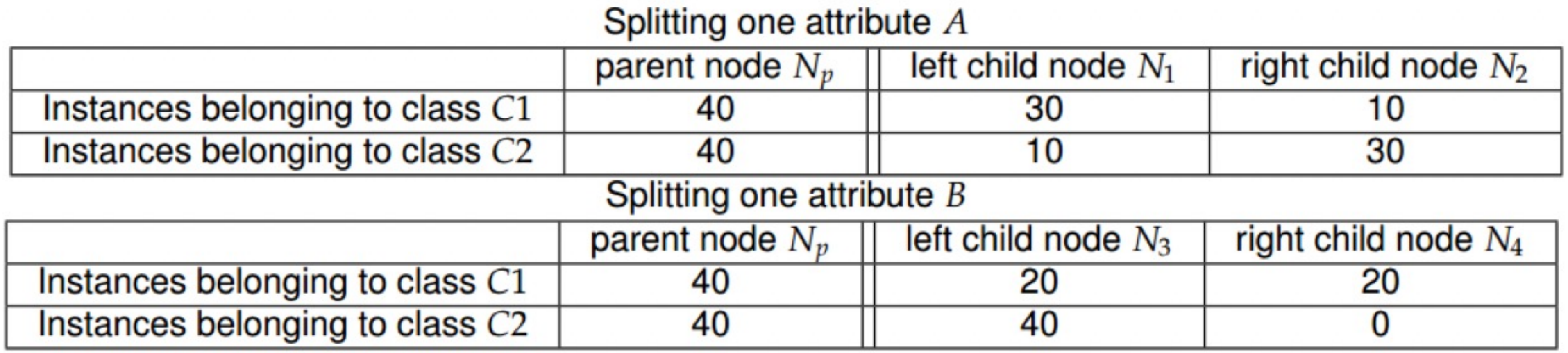
Splitting one attribute A

	parent node N_p	left child node N_1	right child node N_2
Instances belonging to class $C1$	40	30	10
Instances belonging to class $C2$	40	10	30

Splitting one attribute B

	parent node N_p	left child node N_3	right child node N_4
Instances belonging to class $C1$	40	20	20
Instances belonging to class $C2$	40	40	0

Solution steps



Splitting one attribute A

	parent node N_p	left child node N_1	right child node N_2
Instances belonging to class $C1$	40	30	10
Instances belonging to class $C2$	40	10	30

Splitting one attribute B

	parent node N_p	left child node N_3	right child node N_4
Instances belonging to class $C1$	40	20	20
Instances belonging to class $C2$	40	40	0

Using **gini index**:

Splitting on *A*: $\mathrm{IG}\left(\mathrm{N_p}, \mathrm{A}\right) = 0.5 - \frac{4}{8} * 0.375 \; - \frac{4}{8} * 0.375 = 0.125$

Splitting on *B*: $\mathrm{IG}\left(\mathrm{N_p}, \mathrm{B}\right) = 0.5 - \frac{6}{8} * 0.4 \; - 0 = 0.16$

Thus, splitting on *B* is preferred.

Classification – Exercise 6

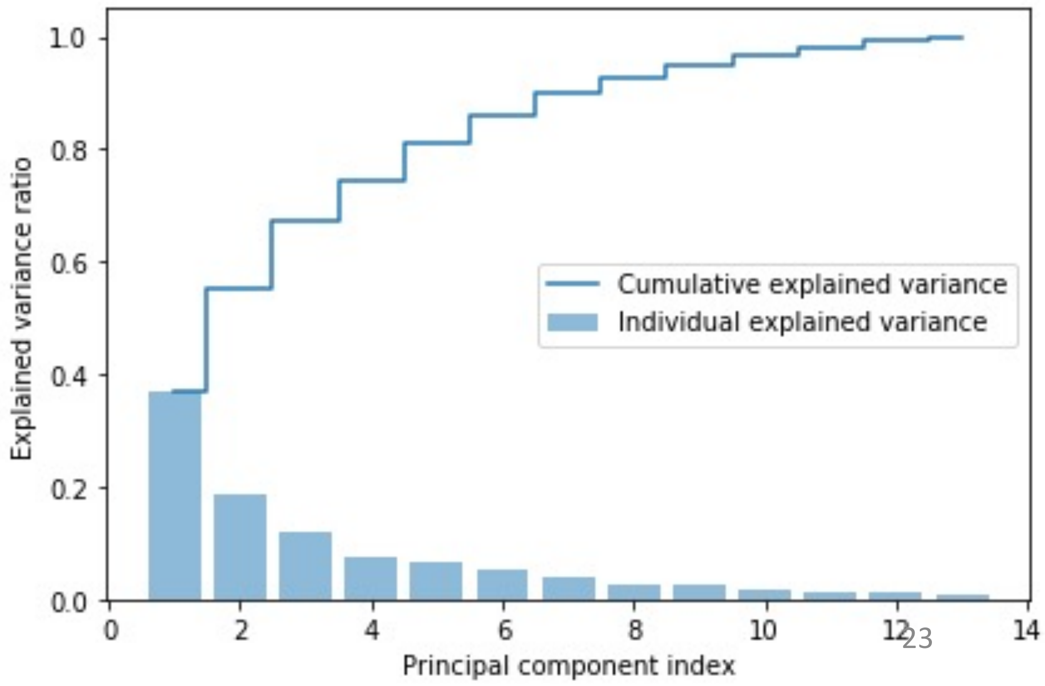
- (T/F) To classify a dataset with 3 class labels, we need to use multiclass classification problem. Using One-versus-Rest (OvR) strategy, we need to build more classifiers than using One-versus-one strategy.
- Which metric cannot be used to evaluate the performance of a classification model?
  - (a) Accuracy
  - (b) Precision
  - (c) F1
  - (d) AUC
  - (e) ROC
  - (f) R^2

Classification – Exercise 7 (Logistic regression)

- What is the output range of the logistic function?
- Let y11=sigmoid(100), y12=sigmoid(200), y21=sigmoid(1) and y22=sigmoid(6). Which value is bigger?
    - |y12-y11|
    - |y22-y21|

Dimension reduction - Exercise 1

- Given the variance-explained-ratios plot for PCA analysis. Assume that we need to choose x number of PCs to account for 60% of the variance. What is the minimum x we should choose?
  - (a) 1
  - (b) 2
  - (c) 3
  - (d) 4

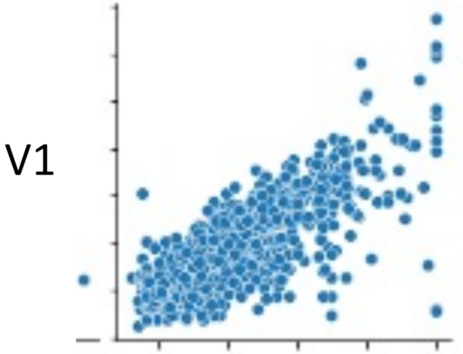


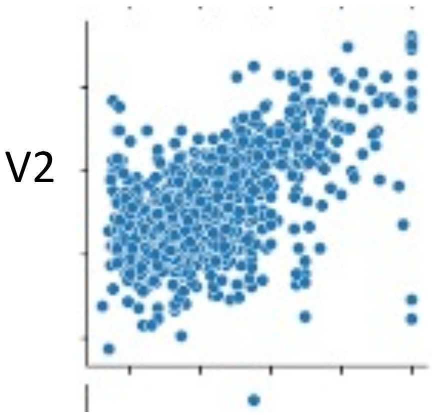
Dimension reduction - Exercise 2

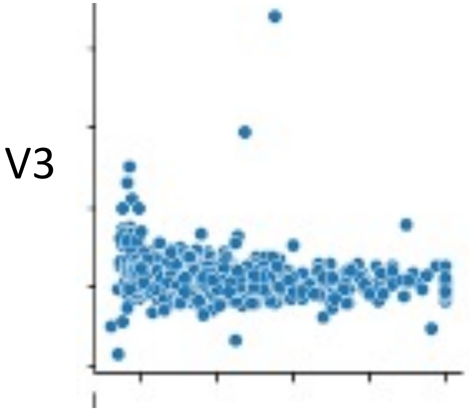
- (T/F) In LDA, the number of linear discriminants is at most $C - 1$ where C is the number of class labels.
- (T/F) In both PCA and LDA analysis, eigen decomposition is used.
- (T/F) LDA can be used as a classifier.

Regression – Exercise 1

- Given the following pair plots, let the x axis represent the same variable V_{target}. Let the y axis of the three plots represent V_1, V_2, and V_3. Which of the variable can better support the prediction of V_{target}? Please explain.



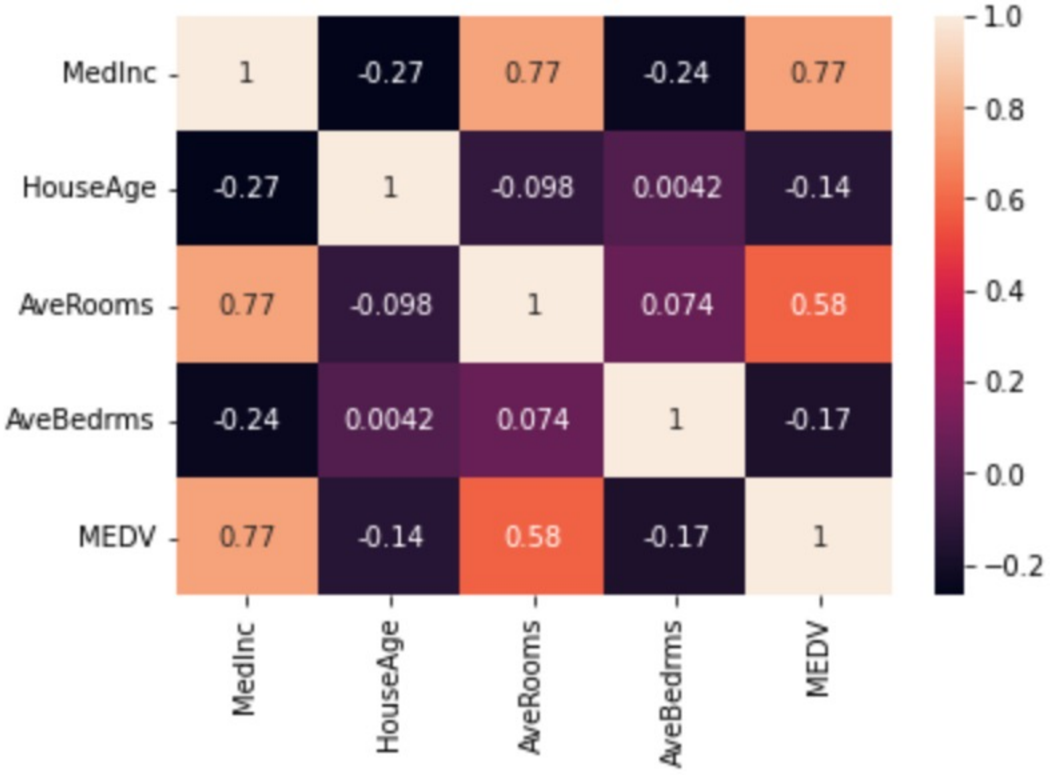




- V1.
- it shows stronger correlation.
- For V3, the changes in V3 does not affect the values of V_{target} much.

Regression - Exercise 2

- Which feature has higher probability to affect the prediction of the target variable MEDV?
    - (a) MedInc
    - (b) HouseAge
    - (c) AveRooms
    - (d) AbeBedrooms



	MedInc	HouseAge	AveRooms	AveBedrms	MEDV
MedInc	1	-0.27	0.77	-0.24	0.77
HouseAge	-0.27	1	-0.098	0.0042	-0.14
AveRooms	0.77	-0.098	1	0.074	0.58
AveBedrms	-0.24	0.0042	0.074	1	-0.17
MEDV	0.77	-0.14	0.58	-0.17	1

Regression - Exercise 3

- What are the differences between the regular linear regression and the RANSAC regressor?
- Given a dataset with two features a and b,
  - how many features will we get if you degree-2 polynomial regression? What are they?
  - How many features will we get if you degree-3 polynomial regression? What are they?
- Which metric cannot be used to evaluate the performance of a regression model?
  - (a) MSE
  - (b) R^2
  - (c) Residual plot
  - (d) Accuracy

Regression – Exercise 4

- (T/F) ElasticNet regression algorithm uses both L1 and L2 regularization.
- (T/F) Ridgge regression model uses L1 regularization.
- (T/F) LASSO regression model uses L2 regularization.
- (T/F) Random forest regressor is a linear regression model.
- (T/F) Decision trees can be used to do regression analysis.
- (T/F) Support vectors can be used to do regression analysis.

Clustering – Exercise 1 (k-means)

Suppose that you are required to cluster the following points into three clusters.
P_1(2,10), P_2(2,5), P_3(8,4), P_4(5,8), P_5(7,5), P_6(6,4), P_7(1,2), P_8(4,9)

The distance function is Euclidean distance. Suppose initially we assign P_1, P_4, and P_{17} as the center of each cluster, respectively. Use the *k*-means algorithm to show

(a) The three cluster centers after the first round of execution.

(b) The final three clusters.

Solution ideas

Iteration 1

center1=P1	2	10
center2=P4	5	8
center3=P7	1	2

Use squared Euclidean distance

$$dist = \ (y2 - y1)^2 + (x2 - x1)^2$$

	feature 1	feature 2	distance to center1	distance to center 2	distance to center 3	Assign to
P1	2	10	0	13	65	C1
P2	2	5	25	18	10	C3
P3	8	4	72	25	53	C2
P4	5	8	13	0	52	C2
P5	7	5	50	13	45	C2
P6	6	4	52	17	29	C2
P7	1	2	65	52	0	C3
P8	4	9	5	2	58	C2

New center 1	2	10
New center 2	6	6
New center 3	1.5	3.5

Solution ideas

- Iteration 2

Center 1	2	10
Center 2	6	6
Center 3	1.5	3.5

	feature 1	feature 2	distance to center1	distance to center 2	distance to center 3	Assign to
P1	2	10	0	32	42.5	C1
P2	2	5	25	17	2.5	C3
P3	8	4	72	8	42.5	C2
P4	5	8	13	5	32.5	C2
P5	7	5	50	2	32.5	C2
P6	6	4	52	4	20.5	C2
P7	1	2	65	41	2.5	C3
P8	4	9	5	13	36.5	C1

New center 1	3	9.5
New center 2	6.5	5.25
New center 3	1.5	3.5

Solution ideas

- Iteration 3

Center 1	3	9.5
Center 2	6.5	5.25
Center 3	1.5	3.5

	feature 1	feature 2	distance to center1	distance to center 2	distance to center 3	Assign to
P1	2	10	1.25	43.7225	42.5	C1
P2	2	5	21.25	21.2225	2.5	C3
P3	8	4	55.25	3.5225	42.5	C2
P4	5	8	6.25	10.1225	32.5	C2
P5	7	5	36.25	0.2225	32.5	C2
P6	6	4	39.25	1.9225	20.5	C2
P7	1	2	60.25	41.9225	2.5	C3
P8	4	9	1.25	20.8225	36.5	C1

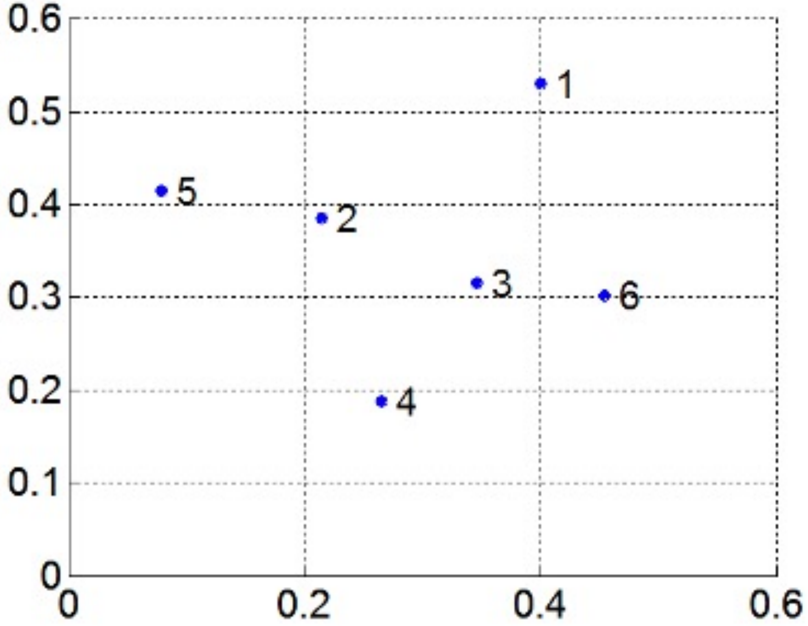
STABLE!

New center 1	3	9.5
New center 2	6.5	5.25
New center 3	1.5	3.5

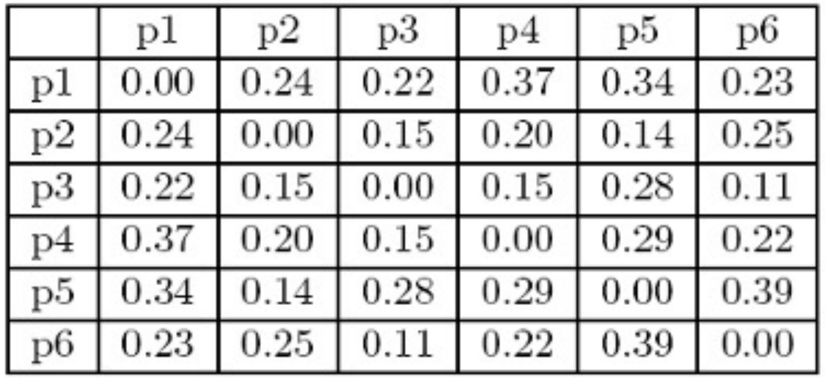
Clustering – Exercise 2 (Hierarchical)

Given the dataset below

- a) show the clustering steps using min/max strategy



Distance Matrix:



	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Ensemble - Exercises

- What is one major difference between the regular bagging approach and random forest method?
- Calculation questions similar to Q1 in HW7.

DNN - Exercises

- Loss functions
  - Definition
  - Where can they be used
  - Advantage, disadvantages
- CNN architecture
  - Convolutional layer (padding, kernel, stride)
  - Pooling layer (kernel, stride)
- RNN architecture