Introduction

Chao Lan

Computer program can automate many tasks...

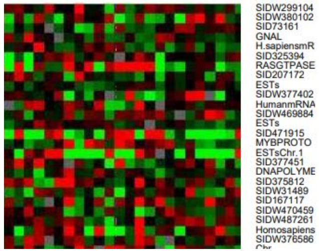
classify image

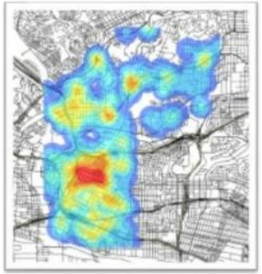
predict student GPA

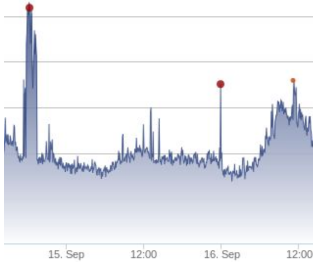
			Lie, Cheat,	Behavior
Student Name	Student ID	Steal	Sneak	Problem
Angel, Julio	2310	0	0	0
Akins, J'Monte	2013	0	0	0
Backer, Brent	2031	0	1	2
Boxwell, Kylie	2001	0	0	0
Cartright, Ashley	2152	0	1	1
Cox, Lucille	2002	0	0	0
Hankins, Erin	2017	0	0	0
Illio, Helen	2132	0	0	0
Jackson, Ronald	2003	0	1	2
Kemp, Patrice	2009	0	0	1
Parker, Stephanie	2004	0	0	0

predict patient condition

predict next purchase







Machine learning helps to build intelligent programs.

Aim to build a computer program that automatically improves its performance through experience.

Example: build a spam filter which automatically improves its filtering accuracy by observing more example emails.

Q: is this a machine learning practice?

Aim to build a computer program that automatically improves its performance through experience.

Example: build a spam filter which automatically improves its filtering accuracy by observing more example emails.

Practice 1: program a spam filter based on three rules

- if an email contains the word “lottery”, it is spam
- if an email contains a link, it is spam
- otherwise, the email is ham

Q: is this a machine learning practice?

Aim to build a computer program that automatically improves its performance through experience.

Example: build a spam filter which automatically improves its filtering accuracy by observing more example emails.

Practice 2: program a spam filter which automatically extracts some “patterns” from a set of given example emails, and uses these patterns to filter spams.

Q: can you describe a (non) machine learning practice?

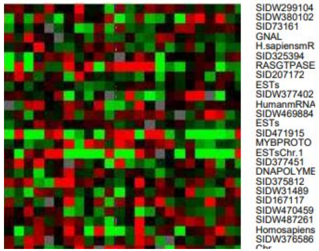
classify image

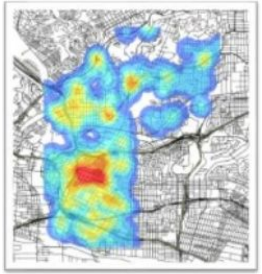
predict student GPA

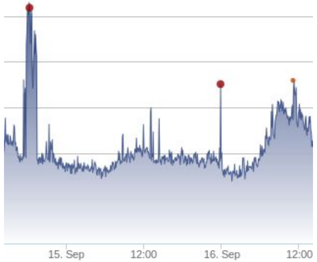
Student Name	Student ID	Steal	Lie, Cheat, Sneak	Behavior Problem
Angel, Julio	2310	0	0	0
Akins, J'Monte	2013	0	0	0
Backer, Brent	2031	0	1	2
Boxwell, Kylie	2001	0	0	0
Cartright, Ashley	2152	0	1	1
Cox, Lucille	2002	0	0	0
Hankins, Erin	2017	0	0	0
Illio, Helen	2132	0	0	0
Jackson, Ronald	2003	0	1	2
Kemp, Patrice	2009	0	0	1
Parker, Stephanie	2004	0	0	0

predict patient condition

predict next purchase





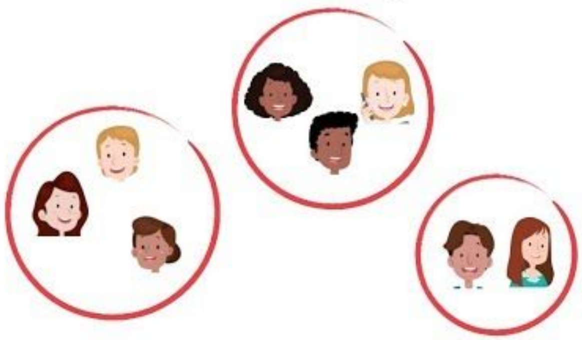


Many Machine Learning Tasks

Prediction

Student Name	Student ID	Steal	Lie, Cheat, Sneak	Behavior Problem	Peer Rejection
Angel, Julio	2310	0	0	0	2
Akins, J'Monte	2013	0	0	0	0
Backer. Brent	2031	0	1	2	1
Boxwell, Kylie	2001	0	0	0	1
Cartright, Ashley	2152	0	1	1	1
Cox, Lucille	2002	0	0	0	0
Hankins, Erin	2017	0	0	0	0
Illio, Helen	2132	0	0	0	0
Jackson, Ronald	2003	0	1	2	2
Kemp, Patrice	2009	0	0	1	0
Parker, Stephanie	2004	0	0	0	0

Clustering



Dimensionality Reduction

Reinforcement Learning

Matrix Recovery

Prediction Task

Three Subjects

Data Representation

Learning Tasks

Model Overfitting

Example Task: Student GPA Prediction

Student Name	Student ID	Steal	Lie, Cheat, Sneak	Behavior Problem	Peer Rejection	Low Academic Achievement	Negative Attitude	Aggressive Behavior	*Total SRSS*	GPA
Angel, Julio	2310	0	0	0	2	0	0	1	**3**	**3.1**
Akins, J'Monte	2013	0	0	0	0	0	0	0	**0**	**4.1**
Backer, Brent	2031	0	1	2	1	2	2	1	**9**	**2.3**
Boxwell, Kylie	2001	0	0	0	1	1	0	0	**2**	**2.5**
Cartright, Ashley	2152	0	1	1	1	0	1	0	**4**	**3.2**
Cox, Lucille	2002	0	0	0	0	0	0	0	**0**	**3.9**
Hankins, Erin	2017	0	0	0	0	0	2	0	**2**	**3.7**
Illio, Helen	2132	0	0	0	0	0	0	0	**0**	**2.9**
Jackson, Ronald	2003	0	1	2	2	3	2	2	**12**	**1.7**
Kemp, Patrice	2009	0	0	1	0	0	0	0	**1**	**3.3**
Parker, Stephanie	2004	0	0	0	0	1	2	0	**3**	**2.7**
Reed, Kent	2010	0	0	0	0	0	0	0	**0**	**3.6**
Sterling, Michael	2022	0	0	1	0	3	1	1	**6**	**2.4**
Thomas, James	2018	0	0	0	0	0	0	0	**0**	**3.8**
Walsh, Carter	2215	0	0	0	1	0	1	0	**2**	**3.5**

Instance, Label

Each student is an instance **x**. GPA is the label **y** (variable to predict).

Student Name	Student ID	Steal	Lie, Cheat, Sneak	Behavior Problem	Peer Rejection	Low Academic Achievement	Negative Attitude	Aggressive Behavior	*Total SRSS*	GPA
Angel, Julio	2310	0	0	0	2	0	0	1	**3**	**3.1**
Akins, J'Monte	2013	0	0	0	0	0	0	0	**0**	**4.1**
Backer, Brent	2031	0	1	2	1	2	2	1	**9**	**2.3**
Boxwell, Kylie	2001	0	0	0	1	1	0	0	**2**	**2.5**
Cartright, Ashley	2152	0	1	1	1	0	1	0	**4**	**3.2**
Cox, Lucille	2002	0	0	0	0	0	0	0	**0**	**3.9**
Hankins, Erin	2017	0	0	0	0	0	2	0	**2**	**3.7**
Illio, Helen	2132	0	0	0	0	0	0	0	**0**	**2.9**
Jackson, Ronald	2003	0	1	2	2	3	2	2	**12**	**1.7**
Kemp, Patrice	2009	0	0	1	0	0	0	0	**1**	**3.3**
Parker, Stephanie	2004	0	0	0	0	1	2	0	**3**	**2.7**
Reed, Kent	2010	0	0	0	0	0	0	0	**0**	**3.6**
Sterling, Michael	2022	0	0	1	0	3	1	1	**6**	**2.4**
Thomas, James	2018	0	0	0	0	0	0	0	**0**	**3.8**
Walsh, Carter	2215	0	0	0	1	0	1	0	**2**	**3.5**

Feature, Feature Vector

Student Name	Student ID	Steal	Lie, Cheat, Sneak	Behavior Problem	Peer Rejection	Low Academic Achievement	Negative Attitude	Aggressive Behavior	*Total SRSS*	GPA
Angel, Julio	2310	0	0	0	2	0	0	1	3	3.1

A feature is a predictive variable.

A feature vector is a vector of features.

$$x = \begin{pmatrix} \text{steal} \\ \text{lie, cheat, sneak} \\ \text{behavior problem} \\ \text{peer rejection} \\ \text{low achievement} \\ \text{negative attitude} \\ \text{aggressive behavior} \\ \text{total srss} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 2 \\ 0 \\ 0 \\ 1 \\ 3 \end{pmatrix}$$

Q: what is the feature dimension?

Student Name	Student ID	Steal	Lie, Cheat, Sneak	Behavior Problem	Peer Rejection	Low Academic Achievement	Negative Attitude	Aggressive Behavior	*Total SRSS*	GPA
Angel, Julio	2310	0	0	0	2	0	0	1	3	3.1

A feature is a predictive variable.

A feature vector is a vector of features.

$$x = \begin{pmatrix} \text{steal} \\ \text{lie, cheat, sneak} \\ \text{behavior problem} \\ \text{peer rejection} \\ \text{low achievement} \\ \text{negative attitude} \\ \text{aggressive behavior} \\ \text{total srss} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 2 \\ 0 \\ 0 \\ 1 \\ 3 \end{pmatrix}$$

Q: can we add “Student ID” as another feature?

Student Name	Student ID	Steal	Lie, Cheat, Sneak	Behavior Problem	Peer Rejection	Low Academic Achievement	Negative Attitude	Aggressive Behavior	*Total SRSS*	GPA
Angel, Julio	2310	0	0	0	2	0	0	1	3	3.1

A feature is a predictive variable.

A feature vector is a vector of features.

$$x = \begin{pmatrix} \text{steal} \\ \text{lie, cheat, sneak} \\ \text{behavior problem} \\ \text{peer rejection} \\ \text{low achievement} \\ \text{negative attitude} \\ \text{aggressive behavior} \\ \text{total srss} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 2 \\ 0 \\ 0 \\ 1 \\ 3 \end{pmatrix}$$

Sample Matrix, Label Vector

Represent a set of instances by a sample matrix **X** and its label vector **Y**.

each column of X is a feature

X =

0	0	0	2	0	0	1	3
0	0	0	0	0	0	0	0
0	1	2	1	2	2	1	9
0	0	0	1	1	0	0	2
0	1	1	1	0	1	0	4
0	0	0	0	0	0	0	0
0	0	0	0	0	2	0	2
0	0	0	0	0	0	0	0
0	1	2	2	3	2	2	12
0	0	1	0	0	0	0	1
0	0	0	0	1	2	0	3
0	0	0	0	0	0	0	0
0	0	1	0	3	1	1	6
0	0	0	0	0	0	0	0
0	0	0	1	0	1	0	2

each row is an instance

label

Y =

3.1
4.1
2.3
2.5
3.2
3.9
3.7
2.9
1.7
3.3
2.7
3.6
2.4
3.8
3.5

X is an *n-by-p* matrix, and Y is an *n-dim* vector.

p features

X =

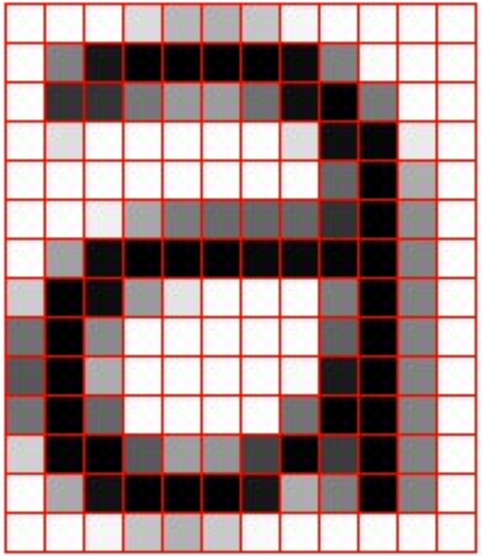
0	0	0	2	0	0	1	3
0	0	0	0	0	0	0	0
0	1	2	1	2	2	1	9
0	0	0	1	1	0	0	2
0	1	1	1	0	1	0	4
0	0	0	0	0	0	0	0
0	0	0	0	0	2	0	2
0	0	0	0	0	0	0	0
0	1	2	2	3	2	2	12
0	0	1	0	0	0	0	1
0	0	0	0	1	2	0	3
0	0	0	0	0	0	0	0
0	0	1	0	3	1	1	6
0	0	0	0	0	0	0	0
0	0	0	1	0	1	0	2

n instances

Y =

3.1
4.1
2.3
2.5
3.2
3.9
3.7
2.9
1.7
3.3
2.7
3.6
2.4
3.8
3.5

Example Feature Vector (Image)



1.0	1.0	1.0	0.9	0.6	0.6	0.6	1.0	1.0	1.0	1.0	1.0
1.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.5	1.0	1.0	1.0
1.0	0.2	0.2	0.5	0.6	0.6	0.5	0.0	0.0	0.5	1.0	1.0
1.0	0.9	1.0	1.0	1.0	1.0	1.0	0.9	0.0	0.0	0.9	1.0
1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.5	0.0	0.5	1.0
1.0	1.0	1.0	0.5	0.5	0.5	0.5	0.5	0.4	0.0	0.5	1.0
1.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	1.0
0.9	0.0	0.0	0.6	1.0	1.0	1.0	1.0	0.5	0.0	0.5	1.0
0.5	0.0	0.6	1.0	1.0	1.0	1.0	1.0	0.5	0.0	0.5	1.0
0.5	0.0	0.7	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.5	1.0
0.6	0.0	0.6	1.0	1.0	1.0	1.0	0.5	0.0	0.0	0.5	1.0
0.9	0.1	0.0	0.6	0.7	0.7	0.5	0.0	0.5	0.0	0.5	1.0
1.0	0.7	0.1	0.0	0.0	0.0	0.1	0.9	0.8	0.0	0.5	1.0
1.0	1.0	1.0	0.8	0.8	0.9	1.0	1.0	1.0	1.0	1.0	1.0

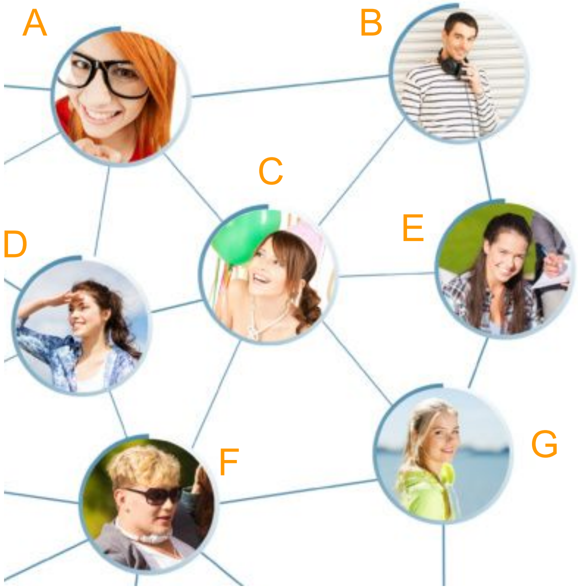
$$x = \begin{pmatrix} 1.0 \\ 1.0 \\ 1.0 \\ 1.0 \\ 1.0 \\ 1.0 \\ 1.0 \\ 0.9 \\ 0.5 \\ 0.5 \\ 0.6 \\ 0.9 \\ 1.0 \\ 1.0 \\ \\ 1.0 \\ 0.5 \\ 0.2 \\ 0.9 \\ 1.0 \\ 1.0 \\ 0.4 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.1 \\ 0.7 \\ 1.0 \\ \cdot \\ \cdot \end{pmatrix}$$

Example Feature Vector (Text)

"It is a puppy and it is extremely cute."

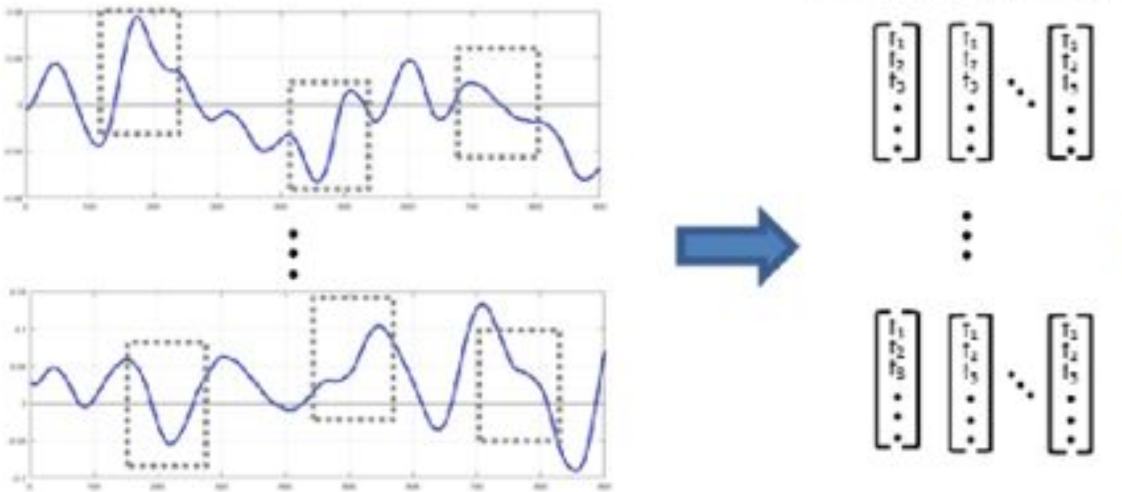
$$x = \begin{pmatrix} \text{it} \\ \text{they} \\ \text{puppy} \\ \text{and} \\ \text{cat} \\ \text{aardvark} \\ \text{cute} \\ \ldots \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ \ldots \end{pmatrix}$$

Example Feature Vector (Network)

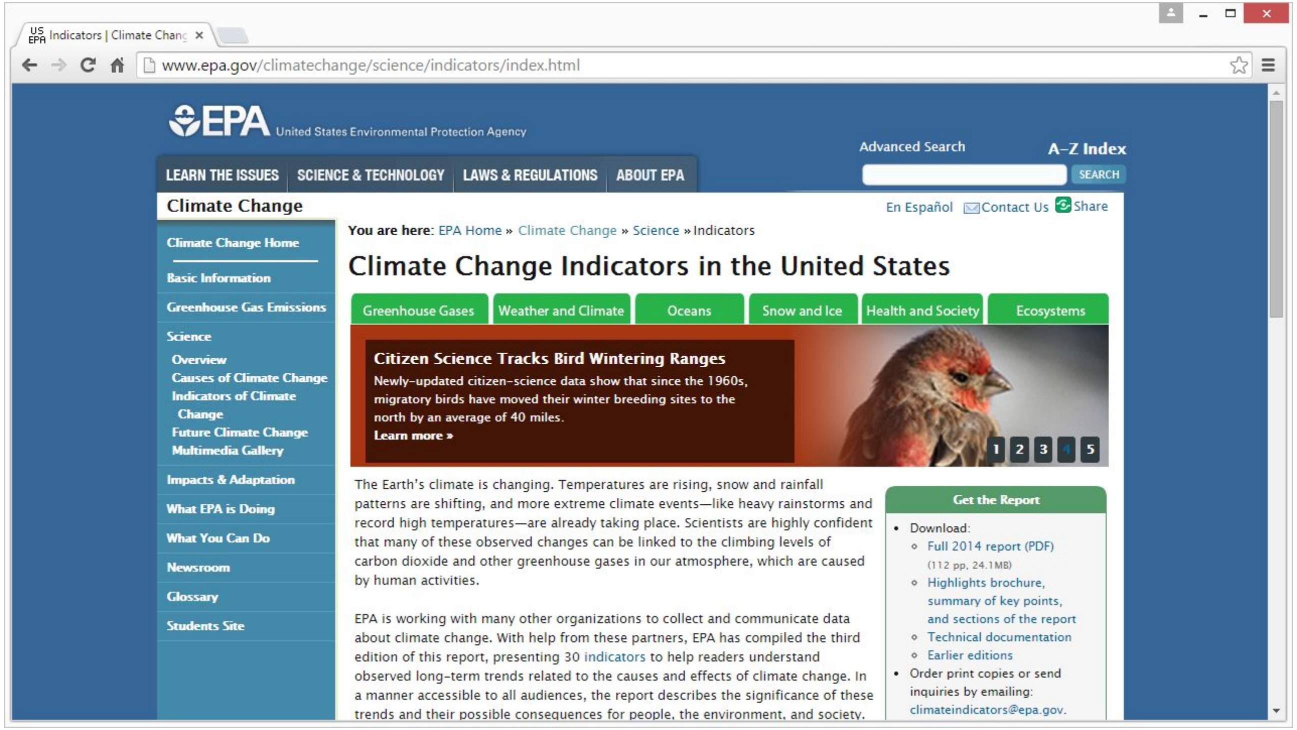


$$x_G = \begin{pmatrix} A? \\ B? \\ C? \\ D? \\ E? \\ F? \\ G? \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

Example Feature Vector (Time-Series/Sequential Data)



Q: how to represent a webpage ? (e.g.,topic prediction)



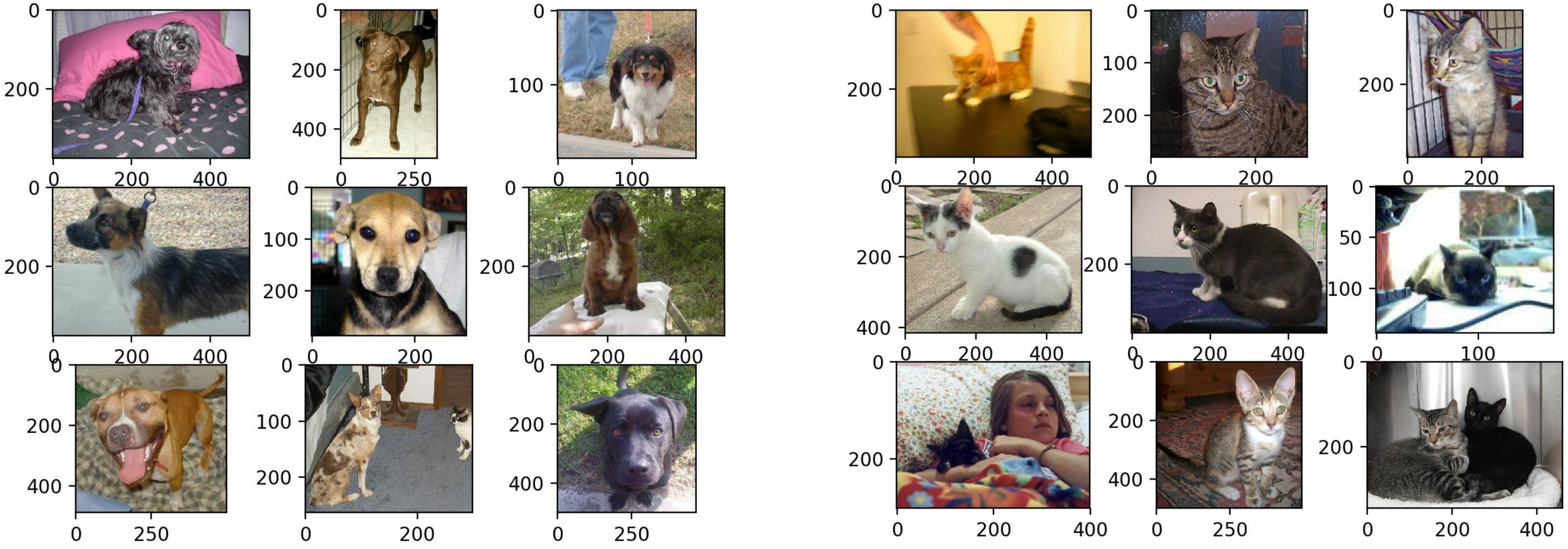
Exercise

Data Representation

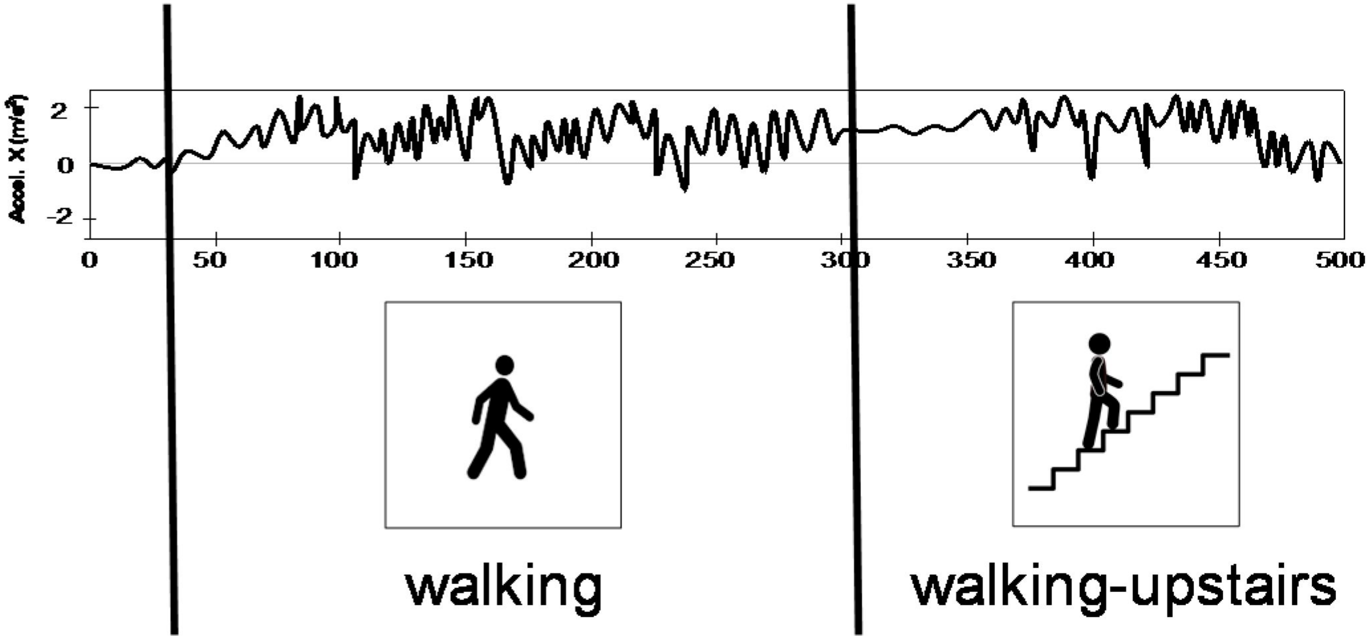
Learning Tasks

Model Overfitting

Q: what are the instance, label and feature in this task?



Q: what are the instance, label and feature in this task?



Three Subjects

Data Representation

Prediction Model

Model Overfitting

Prediction Model, Classification, Regression

A prediction model is a function f. (input x, output predicted y (or, f(x)).

$$x \longrightarrow \text{model } f \longrightarrow f(x)$$

Two Prediction Tasks

- classification: label y is discrete (e.g., predict score as any value in {A,B,C})
- regression: label y is continuous (e.g., predict score as any value in [0,100])

Model Training, Supervised, Unsupervised

Often, f has parameters that need to be learned from a set of training instances, e.g.,

$$f(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_p x_p$$

Two Learning Tasks

- supervised: if training instances are labeled (e.g., a set of students with known GPAs)
- unsupervised: if training instances are not labeled (e.g., same set with unknown GPAs)

In f(x), x_i's are the features of an instance x, and θ_i's are the unknown parameters of f (to be learned).

Hyper-Parameter, Model Complexity

Sometimes, f has hyper-parameters (manually set) that determine its model complexity, e.g.,

$$f(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_p x_p$$

- domain of each θ_i is a hyper-parameter
- larger domain (complex model) can capture more complex relations between x and y

Q: which model *can* be more accurate?

Suppose the true parameters of f(x) are $\theta_0 = -1$, $\theta_1 = 1$, $\theta_2 = 0$, …, $\theta_p = -1$.

$$f(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_p x_p$$

Model 1: f(x) with domain of θ_i set to {0,1}. (smaller domain = simpler model)

Model 2: f(x) with domain of θ_i set to {-2,-1,0,1,2}. (larger domain = more complex model)

Q: so, should we always build a complex model?

Suppose the true parameters of f(x) are $\theta_0 = -1$, $\theta_1 = 1$, $\theta_2 = 0$, …, $\theta_p = -1$.

$$f(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_p x_p$$

Model 1: f(x) with domain of θ_i set to {0,1}. (smaller domain = simpler model)

Model 2: f(x) with domain of θ_i set to {-2,-1,0,1,2}. (larger domain = more complex model)

Q: so, should we always build a complex model?

Suppose the true parameters of f(x) are θ_0= -1, θ_1= 1, θ_2= 0, …, θ_p= -1.

$$f(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_p x_p$$

Model 1: f(x) with domain of θ_i set to {0,1}. (smaller domain = simpler model)

Model 2: f(x) with domain of θ_i set to {-2,-1,0,1,2}. (larger domain = more complex model)

- more complex model needs more data to accurately estimate its parameters (trade-off)

Model Evaluation, Classification Error, MSE

Once f is trained, we can evaluate its prediction error based on two common metrics.

Classification Error (classification task)

- fraction of mis-predicted instances

$$\frac{1}{n}\sum_{i=1}^{n} 1_{\{f(x_i)\neq y_i\}}$$

Mean Squared Error (MSE) (regression task)

- mean squared difference between f(x) and y

$$\frac{1}{n}\sum_{i=1}^{n} [f(x_i) - y_i]^2$$

We have n instances to evaluate error. x_i is the i_{th} instance, y_i is its true label, $f(x_i)$ is its predicted label.

Q: what is the prediction error of f ?

Classification Error

- fraction of mis-predicted instances

Mean Squared Error (MSE)

- mean squared difference between predictions and true labels

$$\frac{1}{n}\sum_{i=1}^{n}[f(x_i) - y_i]^2$$

Task: predict scores of 100 students in {A,B,C}.

- 20 who got A were predicted to get A
- 10 who got A were predicted to get B
- 50 who got B were predicted to get B
- 10 who got C were predicted to get C
- 10 who got C were predicted to get B

Q: what is the prediction error of f ?

Classification Error

- fraction of mis-predicted instances

Mean Squared Error (MSE)

- mean squared difference between predictions and true labels

$$\frac{1}{n}\sum_{i=1}^{n}[f(x_i) - y_i]^2$$

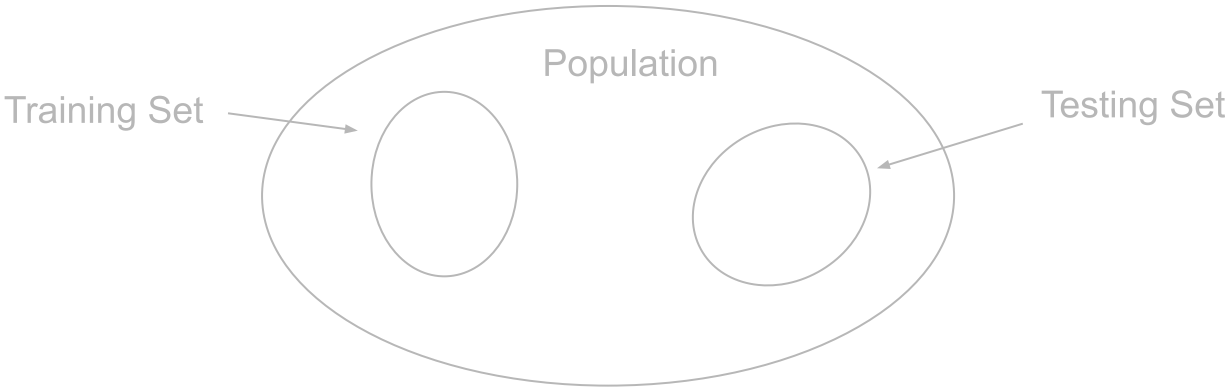
Task: predict GPA of 3 students in [0,4].

- John gets 3.3 and is predicted to get 3.5
- Sam gets 3.0 and is predicted to get 3.0
- Susan gets 3.5 and is predicted to get 3.4

Training/Testing Set, Training/Testing Error

We often use two data sets to evaluate the prediction error of a trained model.

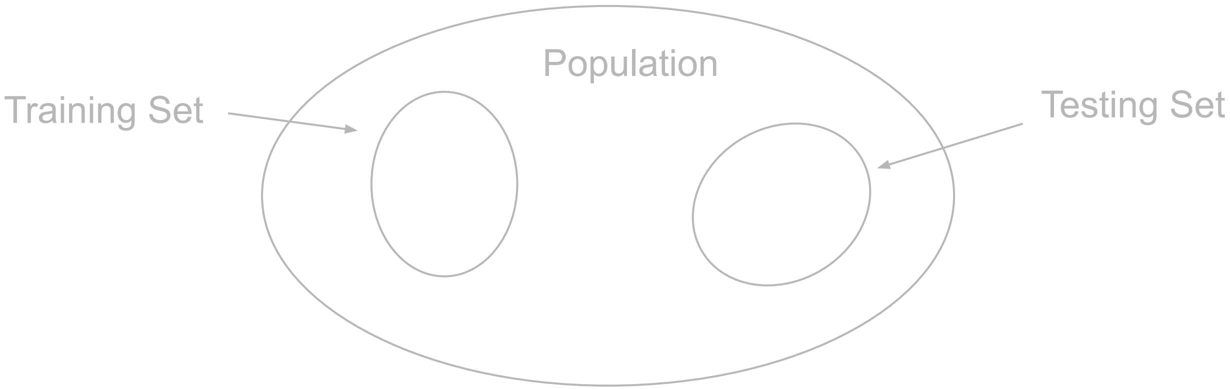
- training set (set of the training instances): error on this set is training error
- testing set (another set of instances): error on this set is testing error



Q: which error better indicates the model performance?

We often use two data sets to evaluate the prediction error of a trained model.

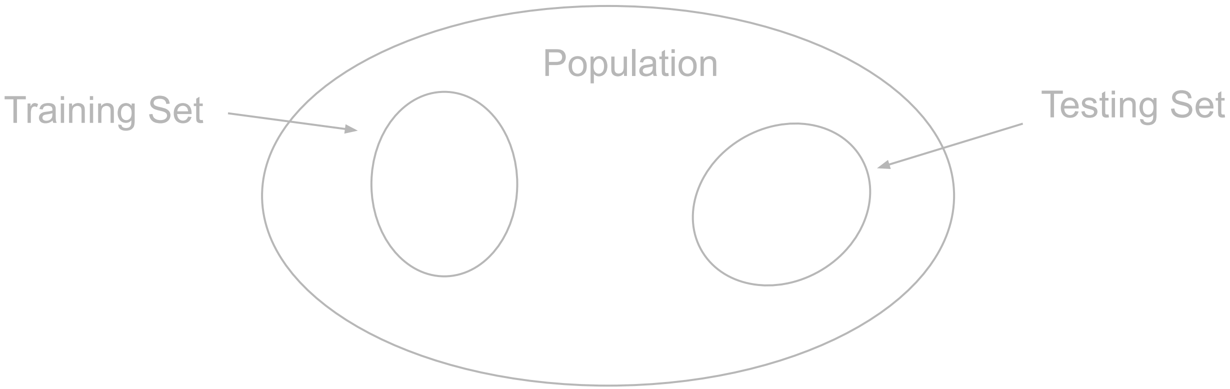
- training set (set of the training instances): error on this set is training error
- testing set (another set of instances): error on this set is testing error



Q: which model do you prefer to use?

Model 1: training error = 0.001, testing error = 0.5

Model 2: training error = 0.1, testing error = 0.2



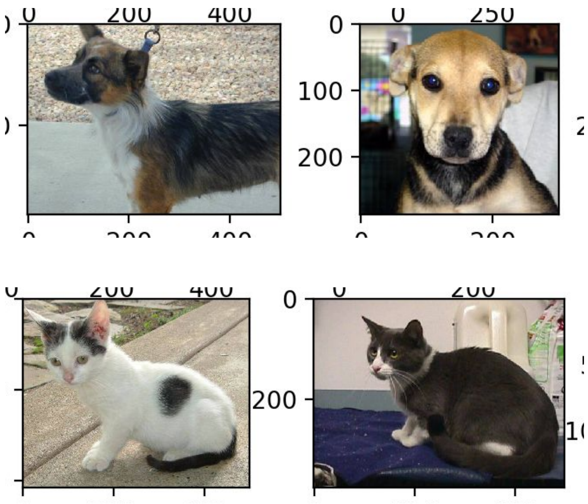
Exercise

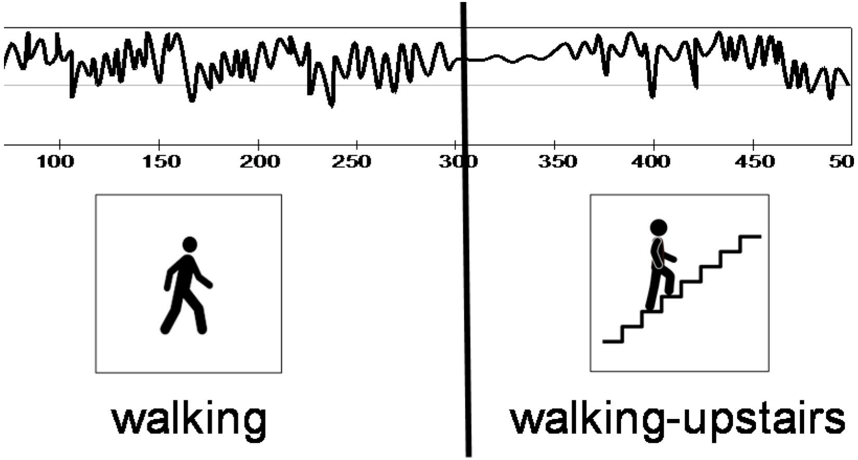
Data Representation

Prediction Model

Model Overfitting

Q: is this a classification or regression task?





Q: how about this task?

Predict Exam Performance

Depend on the grading system (your problem definition)

Grade System	
100	A+
93-99	A
90-92	A-
87-89	B+
83-86	B
80-82	B-
77-79	C+
73-76	C
70-72	C-
67-69	D+
65-66	D
0-64	F

Predict Exam Performance

Q: is this a supervised or unsupervised learning task?

Training Set

Student Name	Student ID	Steal	Lie, Cheat, Sneak	Behavior Problem	Peer Rejection	Low Academic Achievement	Negative Attitude	Aggressive Behavior	*Total SRSS*	GPA
Angel, Julio	2310	0	0	0	2	0	0	1	3	3.1
Akins, J'Monte	2013	0	0	0	0	0	0	0	0	4.1
Backer, Brent	2031	0	1	2	1	2	2	1	9	2.3
Boxwell, Kylie	2001	0	0	0	1	1	0	0	2	2.5
Cartright, Ashley	2152	0	1	1	1	0	1	0	4	3.2
Cox, Lucille	2002	0	0	0	0	0	0	0	0	3.9
Hankins, Erin	2017	0	0	0	0	0	2	0	2	3.7
Illio, Helen	2132	0	0	0	0	0	0	0	0	2.9
Jackson, Ronald	2003	0	1	2	2	3	2	2	12	1.7
Kemp, Patrice	2009	0	0	1	0	0	0	0	1	3.3
Parker, Stephanie	2004	0	0	0	0	1	2	0	3	2.7
Reed, Kent	2010	0	0	0	0	0	0	0	0	3.6
Sterling, Michael	2022	0	0	1	0	3	1	1	6	2.4
Thomas, James	2018	0	0	0	0	0	0	0	0	3.8
Walsh, Carter	2215	0	0	0	1	0	1	0	2	3.5

Q: how about now?

Training Set

Student Name	Student ID	Steal	Lie, Cheat, Sneak	Behavior Problem	Peer Rejection	Low Academic Achievement	Negative Attitude	Aggressive Behavior	*Total SRSS*	GPA
Angel, Julio	2310	0	0	0	2	0	0	1	**3**	
Akins, J'Monte	2013	0	0	0	0	0	0	0	**0**	
Backer, Brent	2031	0	1	2	1	2	2	1	**9**	
Boxwell, Kylie	2001	0	0	0	1	1	0	0	**2**	
Cartright, Ashley	2152	0	1	1	1	0	1	0	**4**	
Cox, Lucille	2002	0	0	0	0	0	0	0	**0**	
Hankins, Erin	2017	0	0	0	0	0	2	0	**2**	
Illio, Helen	2132	0	0	0	0	0	0	0	**0**	
Jackson, Ronald	2003	0	1	2	2	3	2	2	**12**	
Kemp, Patrice	2009	0	0	1	0	0	0	0	**1**	
Parker, Stephanie	2004	0	0	0	0	1	2	0	**3**	
Reed, Kent	2010	0	0	0	0	0	0	0	**0**	
Sterling, Michael	2022	0	0	1	0	3	1	1	**6**	
Thomas, James	2018	0	0	0	0	0	0	0	**0**	
Walsh, Carter	2215	0	0	0	1	0	1	0	**2**	

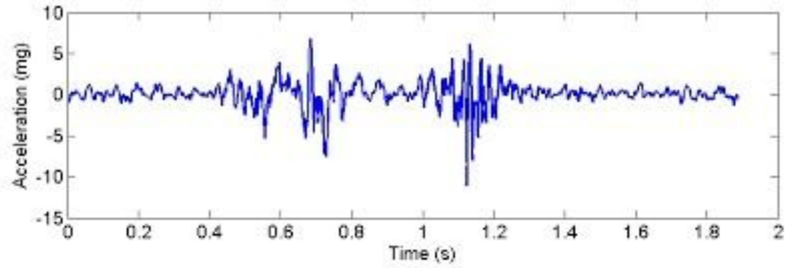
Q: and this?

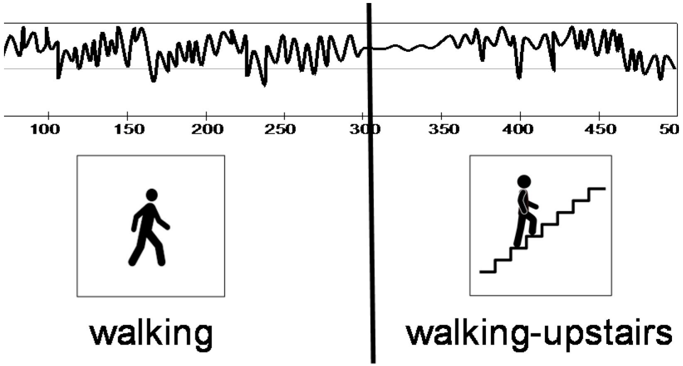
Training Set

Student Name	Student ID	Steal	Lie, Cheat, Sneak	Behavior Problem	Peer Rejection	Low Academic Achievement	Negative Attitude	Aggressive Behavior	*Total SRSS*	GPA
Angel, Julio	2310	0	0	0	2	0	0	1	**3**	
Akins, J'Monte	2013	0	0	0	0	0	0	0	**0**	
Backer, Brent	2031	0	1	2	1	2	2	1	**9**	2.3
Boxwell, Kylie	2001	0	0	0	1	1	0	0	**2**	
Cartright, Ashley	2152	0	1	1	1	0	1	0	**4**	
Cox, Lucille	2002	0	0	0	0	0	0	0	**0**	3.9
Hankins, Erin	2017	0	0	0	0	0	2	0	**2**	
Illio, Helen	2132	0	0	0	0	0	0	0	**0**	
Jackson, Ronald	2003	0	1	2	2	3	2	2	**12**	
Kemp, Patrice	2009	0	0	1	0	0	0	0	**1**	
Parker, Stephanie	2004	0	0	0	0	1	2	0	**3**	
Reed, Kent	2010	0	0	0	0	0	0	0	**0**	
Sterling, Michael	2022	0	0	1	0	3	1	1	**6**	
Thomas, James	2018	0	0	0	0	0	0	0	**0**	
Walsh, Carter	2215	0	0	0	1	0	1	0	**2**	

Q: is this a supervised or unsupervised learning task?

Training Signals





Q: any flaw(s) in the following statements?

I train a model on this set. It gets 0.1 classification error on the set. It is a very good model!

Student Name	Student ID	Steal	Lie, Cheat, Sneak	Behavior Problem	Peer Rejection	Low Academic Achievement	Negative Attitude	Aggressive Behavior	*Total SRSS*	GPA
Angel, Julio	2310	0	0	0	2	0	0	1	3	3.1
Akins, J'Monte	2013	0	0	0	0	0	0	0	0	4.1
Backer, Brent	2031	0	1	2	1	2	2	1	9	2.3
Boxwell, Kylie	2001	0	0	0	1	1	0	0	2	2.5
Cartright, Ashley	2152	0	1	1	1	0	1	0	4	3.2
Cox, Lucille	2002	0	0	0	0	0	0	0	0	3.9
Hankins, Erin	2017	0	0	0	0	0	2	0	2	3.7
Illio, Helen	2132	0	0	0	0	0	0	0	0	2.9
Jackson, Ronald	2003	0	1	2	2	3	2	2	12	1.7
Kemp, Patrice	2009	0	0	1	0	0	0	0	1	3.3
Parker, Stephanie	2004	0	0	0	0	1	2	0	3	2.7
Reed, Kent	2010	0	0	0	0	0	0	0	0	3.6
Sterling, Michael	2022	0	0	1	0	3	1	1	6	2.4
Thomas, James	2018	0	0	0	0	0	0	0	0	3.8
Walsh, Carter	2215	0	0	0	1	0	1	0	2	3.5

Three Subjects

Data Representation

Prediction Model

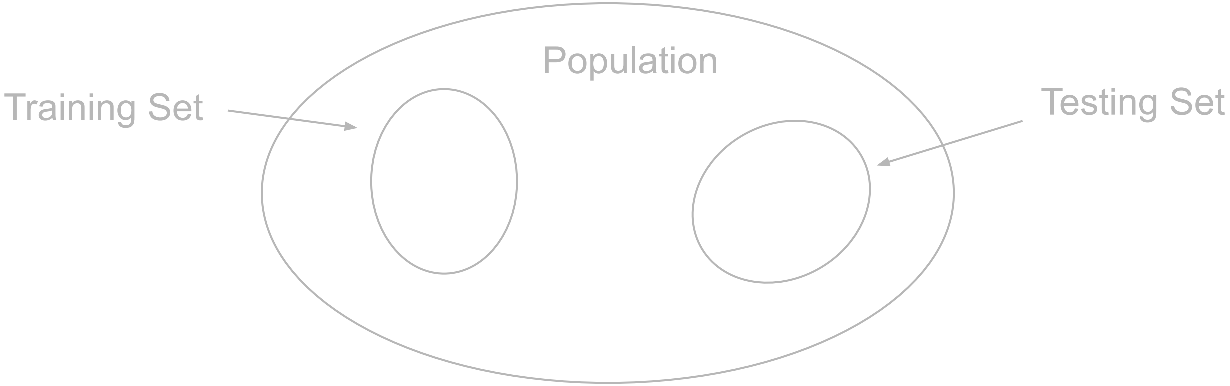
Model Overfitting

Q: which model confuses you most?

Model 1: training error = 0.1, testing error = 0.5

Model 2: training error = 0.1, testing error = 0.2

Model 3: training error = 0.4, testing error = 0.5



Model Overfitting

Overfitting occurs when training error << testing error.

- no strict threshold on the gap

It implies the model fits training data overly well, but does not generalize well on (new) testing data.

It implies the model performance has large variance (more likely to happen on complex model).

Previous Example

M1: train.err = 0.1, test.err = 0.5

M2: train.err = 0.1, test.err = 0.2

M3: train.err = 0.4, test.err = 0.5

Q: what if training error >> testing error?

Overfitting occurs when training error << testing error.

- no strict threshold on the gap

It implies the model fits training data overly well, but does not generalize well on (new) testing data.

It implies the model performance has large variance (more likely to happen on complex model).

Previous Example

M1: train.err = 0.1, test.err = 0.5

M2: train.err = 0.1, test.err = 0.2

M3: train.err = 0.4, test.err = 0.5

M4: train.err = 0.5, test.err = 0.1?

Q: what happens to model 3?

Overfitting occurs when training error << testing error.

- no strict threshold on the gap

It implies the model fits training data overly well, but does not generalize well on (new) testing data.

It implies the model performance has large variance (more likely to happen on complex model).

Previous Example

M1: train.err = 0.1, test.err = 0.5

M2: train.err = 0.1, test.err = 0.2

M3: train.err = 0.4, test.err = 0.5

~~M4: train.err = 0.5, test.err = 0.1~~

Model Underfitting

Underfitting occurs when both errors are similarly big.

- again, no strict threshold

It implies the model cannot fit any data well.

- neither training nor testing, probably not any other

It implies the model performance has large bias and small variance (more likely to occur on simple model).

Previous Example

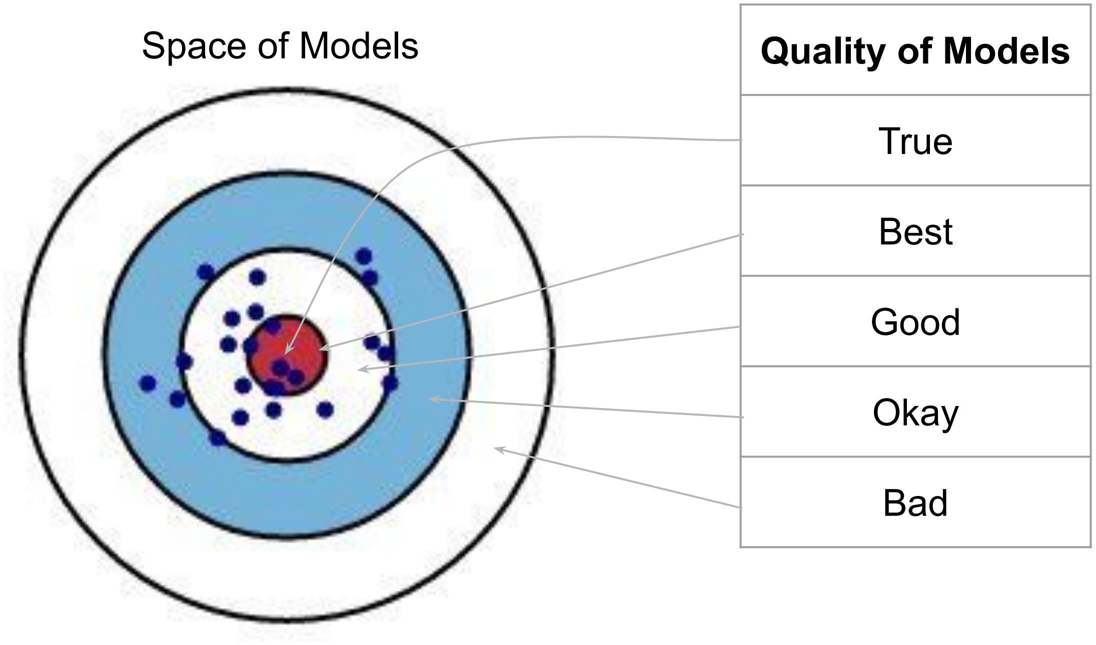
M1: train.err = 0.1,   test.err = 0.5

M2: train.err = 0.1,   test.err = 0.2

M3: train.err = 0.4,   test.err = 0.5

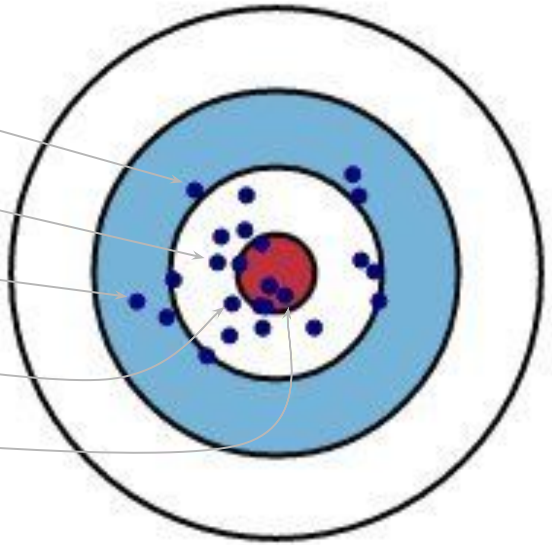
~~M4: train.err = 0.5,   test.err = 0.1~~

Model Variance, Model Bias



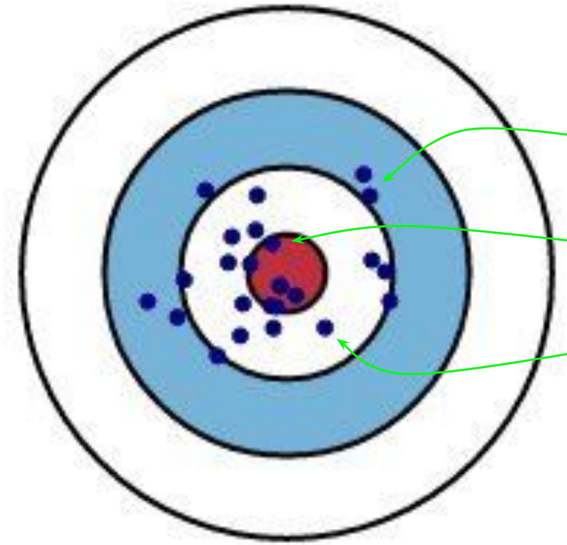
Model Variance, Model Bias

Models trained on
random set 1
random set 2
random set 3
random set 4
random set 5



All sets are randomly drawn from
the same population (assumption).

Model Variance, Model Bias



They have the same function but different parameters (as they learn from different sets).

M1: $f(x) = 0.5 + 0.2\, x_1 - 0.1\, x_2$

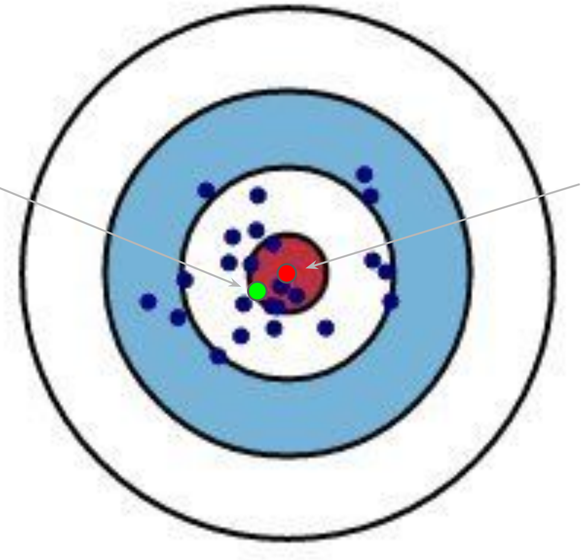
M2: $f(x) = -0.1 + 0.1\, x_1 + 0.4\, x_2$

M3: $f(x) = 0.2 - 0.3\, x_1 + 0.2\, x_2$

Model Variance, Model Bias

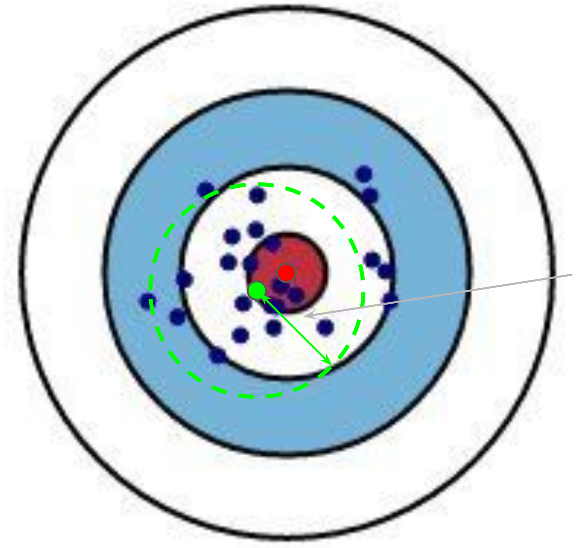
Model Bias

Difference between the true model and averaged model.



We will have more rigorous discussion on bias and variance in later lecture.

Model Variance, Model Bias



Model Variance

Deviation of sample models from the average model.

Indicated by the distance.

Nothing to do with the true model.

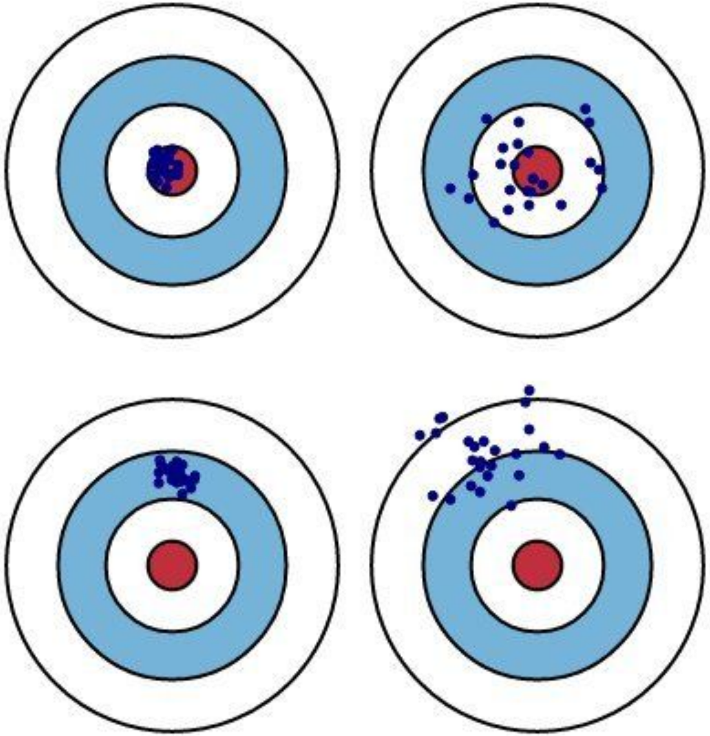
Q: how are their bias and variance? and why?

Model 1

$f(x) = \theta_0 + \theta_1 x_1 + \ldots + \theta_{10} x_{10}$

Model 2

$f(x) = \theta_0 + \theta_1 x_1 + \ldots + \theta_{50} x_{50}$



Model 3

$f(x) = \theta_0 + \theta_1 x_1 + \ldots + \theta_5 x_5$

Model 4

$f(x) = \theta_1 x_1 + \ldots + \theta_{50} x_{50}$

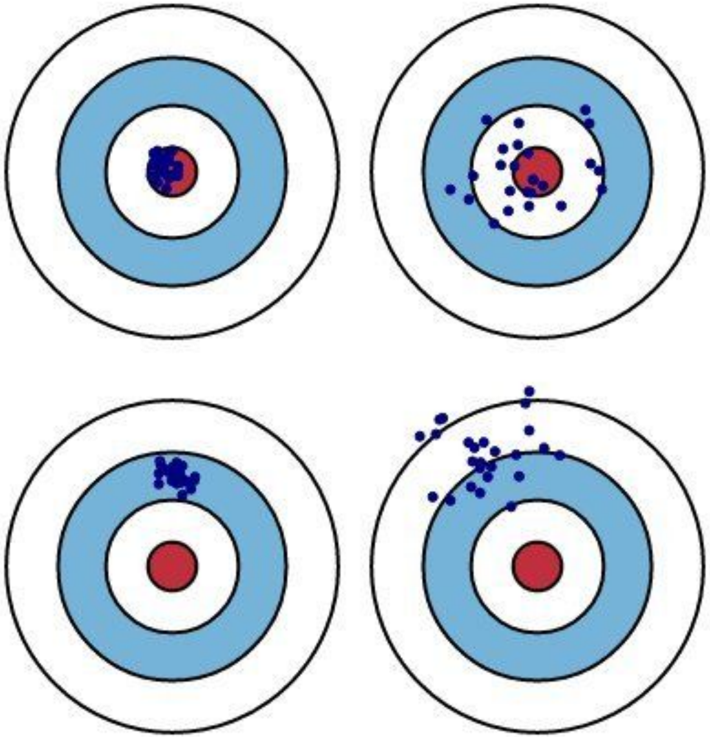
Q: any sign of overfitting or underfitting?

Model 1

$f(x) = \theta_0 + \theta_1 x_1 + \ldots + \theta_{10} x_{10}$

Model 2

$f(x) = \theta_0 + \theta_1 x_1 + \ldots + \theta_{50} x_{50}$



Model 3

$f(x) = \theta_0 + \theta_1 x_1 + \ldots + \theta_5 x_5$

Model 4

$f(x) = \theta_1 x_1 + \ldots + \theta_{50} x_{50}$

Relation: Model Complexity, Overfitting, Variance, Bias

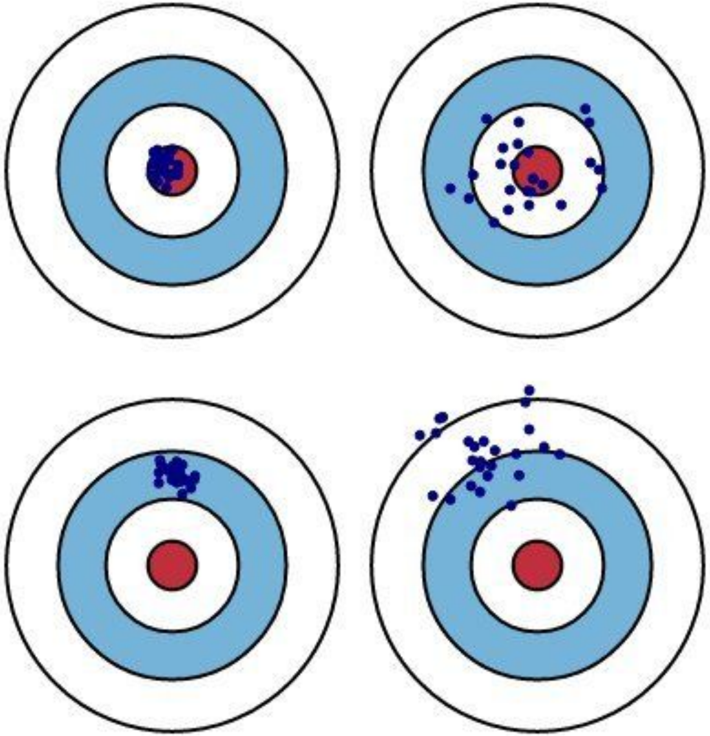
Model 1

$f(x) = \theta_0 + \theta_1 x_1 + \ldots + \theta_{10} x_{10}$

complex model, overfit

large var, small bias

$f(x) = \theta_0 + \theta_1 x_1 + \ldots + \theta_{50} x_{50}$



simple model, underfit

small var, large bias

$f(x) = \theta_0 + \theta_1 x_1 + \ldots + \theta_5 x_5$

Model 4

$f(x) = \theta_1 x_1 + \ldots + \theta_{50} x_{50}$

Tips to avoid learning an overfitted model.

Data

Increase training data.

This helps to reduce model variance without increasing model bias.

Model

Pick a simpler model.

The new model will have smaller variance, probably larger bias.

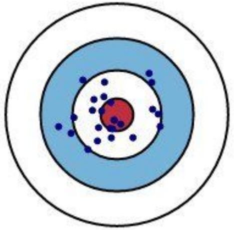
Learner

Apply regularization.

Reduce complexity of a given model by shrinking its parameter domains. Smaller var, probably larger bias.

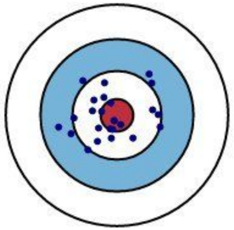
Example

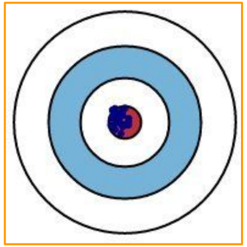
Current Model	Domain of θ
$f(x) = \theta_0 + \theta_1 x_1 + \ldots + \theta_{50} x_{50}$	[0, 10]
Pick a simpler model	Domain of θ
$f(x) = \theta_0 + \theta_1 x_1 + \ldots + \theta_{10} x_{10}$	[0, 10]
Regularized model	Domain of θ
$f(x) = \theta_0 + \theta_1 x_1 + \ldots + \theta_{50} x_{50}$	[0, 5]



Q: which model is most likely to generate which result?

Current Model $f(x) = \theta_0 + \theta_1 x_1 + \ldots + \theta_{50} x_{50}$	Domain of θ [0, 10]
Pick a simpler model $f(x) = \theta_0 + \theta_1 x_1 + \ldots + \theta_{10} x_{10}$	Domain of θ [0, 10]
Regularized model $f(x) = \theta_0 + \theta_1 x_1 + \ldots + \theta_{50} x_{50}$	Domain of θ [0, 5]





More on regularization.

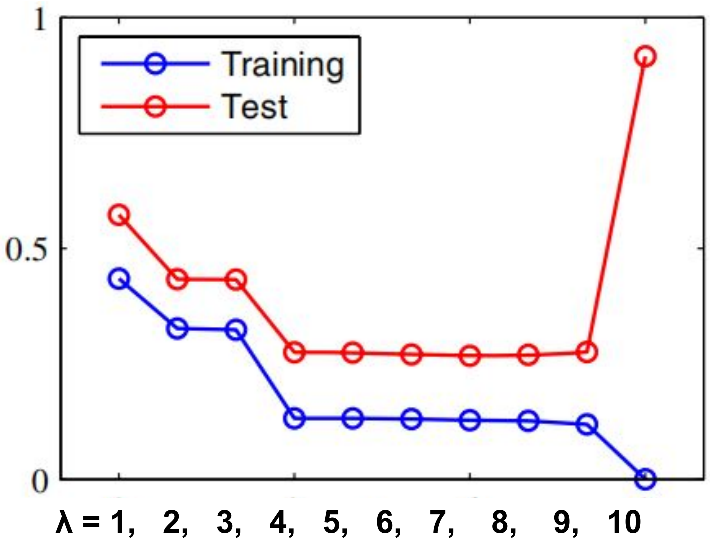
Regularization degree is often controlled by hyperparameter(s).

E.g., domain of θ is [0, λ], where λ is a hyperparameter.

- smaller λ = smaller domain

  = lower model complexity

  = less overfitting



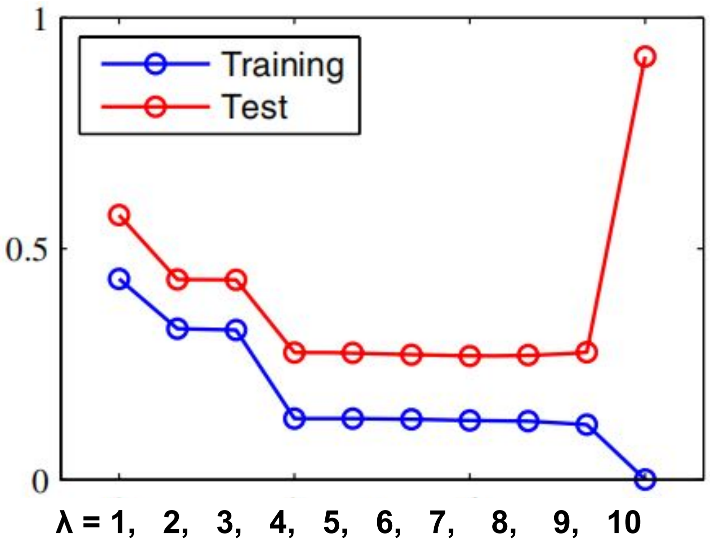
More on regularization.

Regularization degree is often controlled by hyperparameter(s).

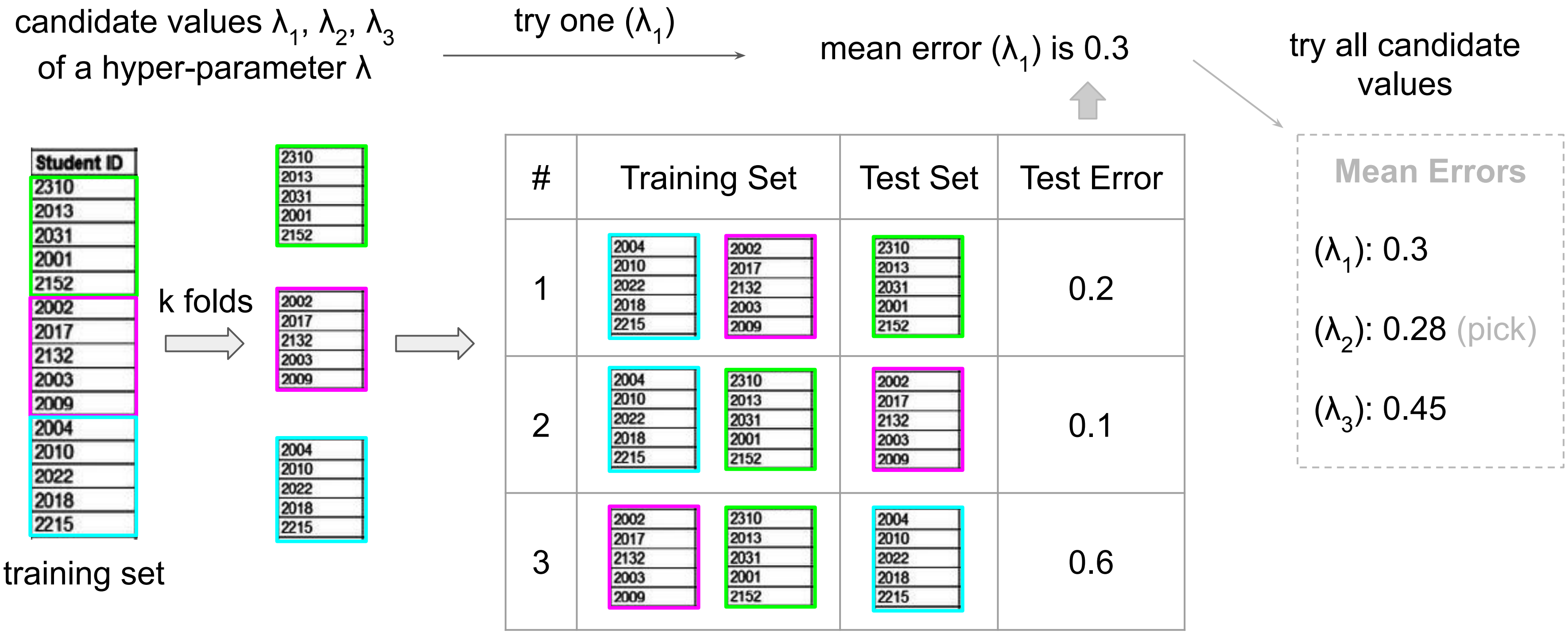
E.g., domain of θ is [0, λ], where λ is a hyperparameter.

We can choose an optimal value for λ by

- experience (e.g., based on the right figure)
- k-fold cross validation



K-Fold Cross Validation (example: K=3)



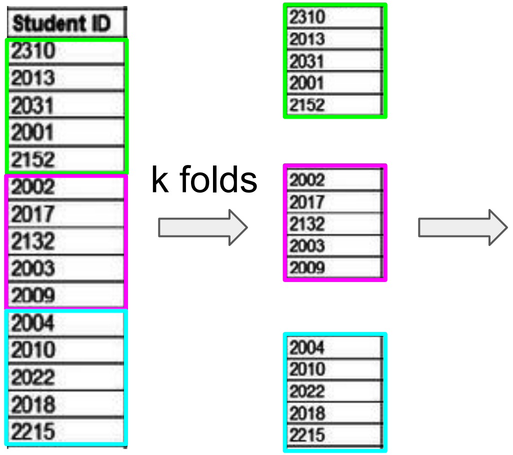
Q: how does CV help to avoid overfitted model?

candidate values λ_1, λ_2, λ_3 of a hyper-parameter λ

try one (λ_1)

mean error (λ_1) is 0.3

try all candidate values



training set

#	Training Set	Test Set	Test Error
1	2004, 2010, 2022, 2018, 2215; 2002, 2017, 2132, 2003, 2009	2310, 2013, 2031, 2001, 2152	0.2
2	2004, 2010, 2022, 2018, 2215; 2310, 2013, 2031, 2001, 2152	2002, 2017, 2132, 2003, 2009	0.1
3	2002, 2017, 2132, 2003, 2009; 2310, 2013, 2031, 2001, 2152	2004, 2010, 2022, 2018, 2215	0.6

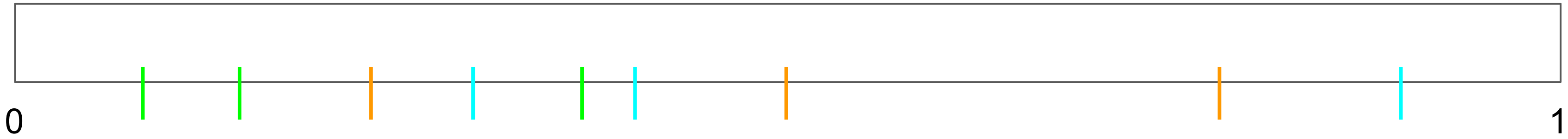
Mean Errors

(λ_1): 0.3

(λ_2): 0.28 (pick)

(λ_3): 0.45

Ways to generate candidate values λ_1, λ_2, λ_3.



manual

0.1, 0.2, 0.35

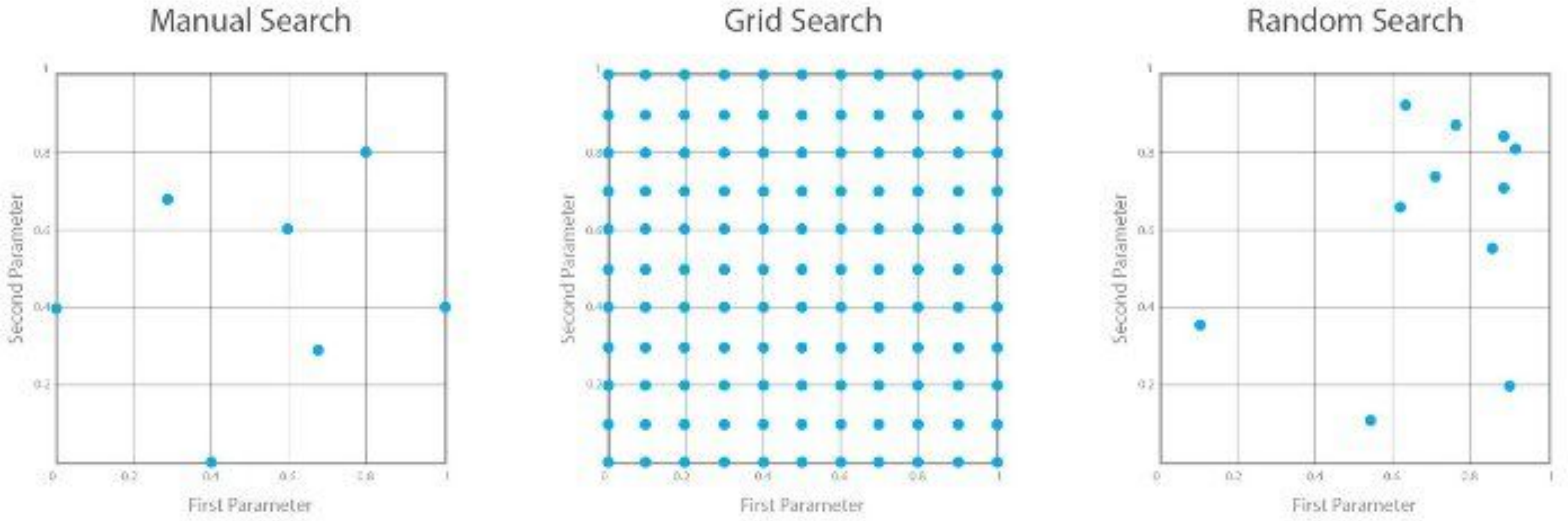
grid search

0.25, 0.5, 0.75

random

0.3, 0.4, 0.9

Ways to generate candidates for two hyperparameters.



Exercise

Data Representation

Prediction Model

Model Overfitting

Q: can you help Susan?

My model just got an error of 0.5!

This is an unacceptable result...

You: is that training error or testing error?

Well, it is training error.

I probably don't have enough training data.

You: is that training error or testing error?

~~Well, it is training error.~~

~~I probably don't have enough training data.~~

Well, it is training error.

I should probably look at testing error instead.

Q: what is likely to happen here? how can she fix it?

Well, it is training error.

~~I probably don't have enough training data.~~

Well, it is training error.

~~I should probably look at testing error instead.~~

You: is that training error or testing error? (cont)

Well, it is testing error. It must be overfitting.

You: what is your training error?

Well, it is testing error. It must be overfitting.

My training error is 0.4. It's overfitting, right?

You: what is your training error?

Well, it is testing error. It must be overfitting.

~~My training error is 0.4. It's overfitting, right?~~

My training error is 0.1. It's overfitting, right?

Q: is she addressing overfitting properly?

ok...so it is overfitting.

I should pick a simpler model, right?

(Revisit) Tips to avoid learning an overfitted model.

Data

Increase training data.

This helps to reduce model variance without increasing model bias.

Model

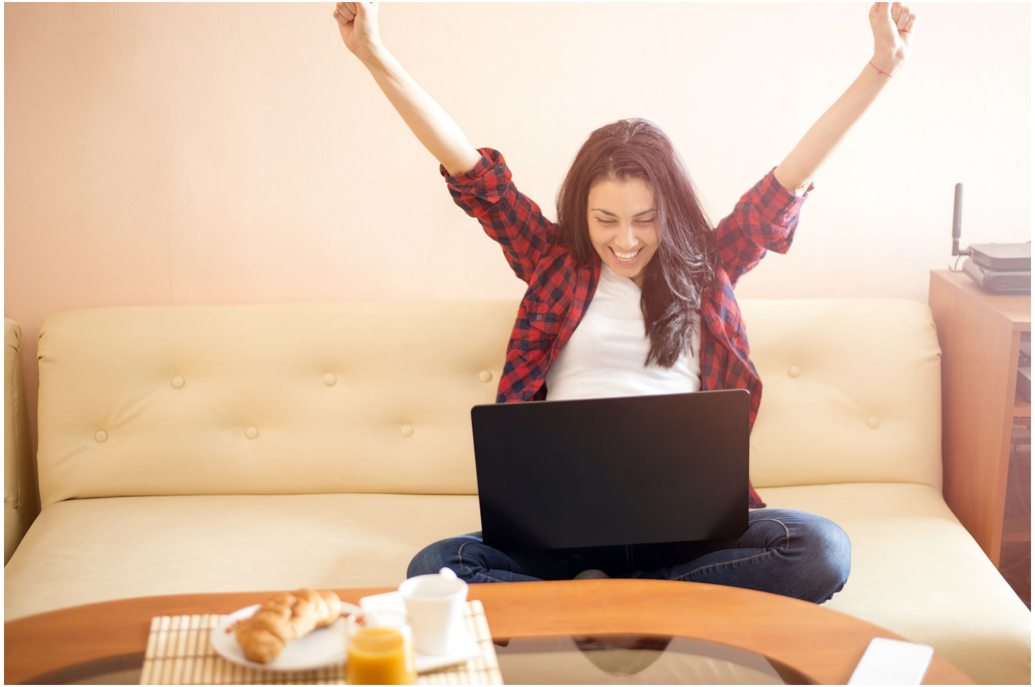
Pick a simpler model.

The new model will have smaller variance, probably larger bias.

Learner

Apply regularization.

Reduce complexity of a given model by shrinking its parameter domains. Smaller var, probably larger bias.



Building a Prediction Model
in Practice

Given a data set, build a model and tell me its performance.

Student Name	Student ID	Steal	Lie, Cheat, Sneak	Behavior Problem	Peer Rejection	Low Academic Achievement	Negative Attitude	Aggressive Behavior	*Total SRSS*	GPA
Angel, Julio	2310	0	0	0	2	0	0	1	**3**	**3.1**
Akins, J'Monte	2013	0	0	0	0	0	0	0	**0**	**4.1**
Backer, Brent	2031	0	1	2	1	2	2	1	**9**	**2.3**
Boxwell, Kylie	2001	0	0	0	1	1	0	0	**2**	**2.5**
Cartright, Ashley	2152	0	1	1	1	0	1	0	**4**	**3.2**
Cox, Lucille	2002	0	0	0	0	0	0	0	**0**	**3.9**
Hankins, Erin	2017	0	0	0	0	0	2	0	**2**	**3.7**
Illio, Helen	2132	0	0	0	0	0	0	0	**0**	**2.9**
Jackson, Ronald	2003	0	1	2	2	3	2	2	**12**	**1.7**
Kemp, Patrice	2009	0	0	1	0	0	0	0	**1**	**3.3**
Parker, Stephanie	2004	0	0	0	0	1	2	0	**3**	**2.7**
Reed, Kent	2010	0	0	0	0	0	0	0	**0**	**3.6**
Sterling, Michael	2022	0	0	1	0	3	1	1	**6**	**2.4**
Thomas, James	2018	0	0	0	0	0	0	0	**0**	**3.8**
Walsh, Carter	2215	0	0	0	1	0	1	0	**2**	**3.5**

1. Examine the data set (feature, label, missing value, etc).

Student Name	Student ID	Steal	Lie, Cheat, Sneak	Behavior Problem	Peer Rejection	Low Academic Achievement	Negative Attitude	Aggressive Behavior	*Total SRSS*	GPA
Angel, Julio	2310	0	0	0	2	0	0	1	**3**	**3.1**
Akins, J'Monte	2013	0	0	0	0	0	0	0	**0**	**4.1**
Backer, Brent	2031	0	1	2	1	2	2	1	**9**	**2.3**
Boxwell, Kylie	2001	0	0	0	1	1	0	0	**2**	**2.5**
Cartright, Ashley	2152	0	1	1	1	0	1	0	**4**	**3.2**
Cox, Lucille	2002	0	0	0	0	0	0	0	**0**	**3.9**
Hankins, Erin	2017	0	0	0	0	0	2	0	**2**	**3.7**
Illio, Helen	2132	0	0	0	0	0	0	0	**0**	**2.9**
Jackson, Ronald	2003	0	1	2	2	3	2	2	**12**	**1.7**
Kemp, Patrice	2009	0	0	1	0	0	0	0	**1**	**3.3**
Parker, Stephanie	2004	0	0	0	0	1	2	0	**3**	**2.7**
Reed, Kent	2010	0	0	0	0	0	0	0	**0**	**3.6**
Sterling, Michael	2022	0	0	1	0	3	1	1	**6**	**2.4**
Thomas, James	2018	0	0	0	0	0	0	0	**0**	**3.8**
Walsh, Carter	2215	0	0	0	1	0	1	0	**2**	**3.5**

Q: what is a simple way to deal with these missing values?

Student Name	Student ID	Steal	Lie, Cheat, Sneak	Behavior Problem	Peer Rejection	Low Academic Achievement	Negative Attitude	Aggressive Behavior	*Total SRSS*	GPA
Angel, Julio	2310	0	0	0	2	0	0	1	3	3.1
Akins, J'Monte	2013	0	0	0	0	0	0	0	0	4.1
Backer, Brent	2031	0	1	2	1	2	2	1	9	2.3
Boxwell, Kylie	2001	0	0	0	1	1	0	0	2	2.5
Cartright, Ashley	2152	0	1	1	1	0	1	0	4	3.2
Cox, Lucille	2002	0	??	0	0	0	0	0	0	3.9
Hankins, Erin	2017	0	0	0	0	0	2	0	2	3.7
Illio, Helen	2132	0	0	0	0	0	0	0	0	2.9
Jackson, Ronald	2003	0	1	2	2	3	2	2	12	1.7
Kemp, Patrice	2009	0	0	1	0	0	0	0	1	3.3
Parker, Stephanie	2004	0	0	0	0	1	2	0	3	2.7
Reed, Kent	2010	0	??	0	0	0	0	0	0	3.6
Sterling, Michael	2022	0	0	1	0	3	1	1	6	2.4
Thomas, James	2018	0	0	0	0	0	0	0	0	3.8
Walsh, Carter	2215	0	0	0	1	0	1	0	2	3.5

Ignore incomplete instance, or apply imputation.

Student Name	Student ID	Steal	Lie, Cheat, Sneak	otal RSS	GPA
Angel, Julio	2310	0	0	3	3.1
Akins, J'Monte	2013	0	0	0	4.1
Backer, Brent	2031	0	1	9	2.3
Boxwell, Kylie	2001	0	0	2	2.5
Cartright, Ashley	2152	0	1	4	3.2
Cox, Lucille	2002	0	??	0	3.9
Hankins, Erin	2017	0	0	2	3.7
Illio, Helen	2132	0	0	0	2.9
Jackson, Ronald	2003	0	1	12	1.7
Kemp, Patrice	2009	0	0	1	3.3
Parker, Stephanie	2004	0	0	3	2.7
Reed, Kent	2010	0	??	0	3.6
Sterling, Michael	2022	0	0	6	2.4
Thomas, James	2018	0	0	0	3.8
Walsh, Carter	2215	0	0	2	3.5

Mean Imputation

Fill the missing value of a feature by the mean of its observed values.

e.g., fill ?? with (0+0+1+0+...+0+0)/13.

Q: what about these missing values?

Student Name	Student ID	Steal	Lie, Cheat, Sneak	Behavior Problem	Peer Rejection	Low Academic Achievement	Negative Attitude	Aggressive Behavior	*Total SRSS*	GPA
Angel, Julio	2310	0	0	0	2	0	0	1	3	3.1
Akins, J'Monte	2013	0	0	0	??	0	0	0	0	4.1
Backer, Brent	2031	0	1	2	??	2	2	1	9	2.3
Boxwell, Kylie	2001	0	0	0	1	1	0	0	2	2.5
Cartright, Ashley	2152	0	1	1	??	0	1	0	4	3.2
Cox, Lucille	2002	0	0	0	??	0	0	0	0	3.9
Hankins, Erin	2017	0	0	0	0	0	2	0	2	3.7
Illio, Helen	2132	0	0	0	??	0	0	0	0	2.9
Jackson, Ronald	2003	0	1	2	??	3	2	2	12	1.7
Kemp, Patrice	2009	0	0	1	??	0	0	0	1	3.3
Parker, Stephanie	2004	0	0	0	??	1	2	0	3	2.7
Reed, Kent	2010	0	0	0	0	0	0	0	0	3.6
Sterling, Michael	2022	0	0	1	??	3	1	1	6	2.4
Thomas, James	2018	0	0	0	??	0	0	0	0	3.8
Walsh, Carter	2215	0	0	0	1	0	1	0	2	3.5

2. Split data into training set (75%) and testing set (25%).

Student Name	Student ID	Steal	Lie, Cheat, Sneak	Behavior Problem	Peer Rejection	Low Academic Achievement	Negative Attitude	Aggressive Behavior	*Total SRSS*	GPA
Angel, Julio	2310	0	0	0	2	0	0	1	3	3.1
Akins, J'Monte	2013	0	0	0	0	0	0	0	0	4.1
Backer, Brent	2031	0	1	2	1	2	2	1	9	2.3
Boxwell, Kylie	2001	0	0	0	1	1	0	0	2	2.5
Cartright, Ashley	2152	0	1	1	1	0	1	0	4	3.2
Cox, Lucille	2002	0	0	0	0	0	0	0	0	3.9
Hankins, Erin	2017	0	0	0	0	0	2	0	2	3.7
Illio, Helen	2132	0	0	0	0	0	0	0	0	2.9
Jackson, Ronald	2003	0	1	2	2	3	2	2	12	1.7
Kemp, Patrice	2009	0	0	1	0	0	0	0	1	3.3
Parker, Stephanie	2004	0	0	0	0	1	2	0	3	2.7
Reed, Kent	2010	0	0	0	0	0	0	0	0	3.6
Sterling, Michael	2022	0	0	1	0	3	1	1	6	2.4
Thomas, James	2018	0	0	0	0	0	0	0	0	3.8
Walsh, Carter	2215	0	0	0	1	0	1	0	2	3.5

3. Set hyper-parameters (experience/CV on *training set*)

Student Name	Student ID	Steal	Lie, Cheat, Sneak	Behavior Problem	Peer Rejection	Low Academic Achievement	Negative Attitude	Aggressive Behavior	*Total SRSS*	GPA
Angel, Julio	2310	0	0	0	2	0	0	1	3	3.1
Akins, J'Monte	2013	0	0	0	0	0	0	0	0	4.1
Backer, Brent	2031	0	1	2	1	2	2	1	9	2.3
Boxwell, Kylie	2001	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	2	2.5
Cartright, Ashley	2152	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	4	3.2
Cox, Lucille	2002	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	0	3.9
Hankins, Erin	2017	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	2	3.7
Illio, Helen	2132	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	0	2.9
Jackson, Ronald	2003	0	1	2	2	3	2	2	12	1.7
Kemp, Patrice	2009	0	0	1	0	0	0	0	1	3.3
Parker, Stephanie	2004	0	0	0	0	1	2	0	3	2.7
Reed, Kent	2010	0	0	0	0	0	0	0	0	3.6
Sterling, Michael	2022	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	6	2.4
Thomas, James	2018	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	3.8
Walsh, Carter	2215	0	0	0	1	0	1	0	2	3.5

Divide training set into K folds, and apply k-fold cross-validation to set hyperparameters.

Testing set is NOT used to determine hyperparameter.

Q: why not use testing set to determine hyperparameter?

Student Name	Student ID	Steal	Lie, Cheat, Sneak	Behavior Problem	Peer Rejection	Low Academic Achievement	Negative Attitude	Aggressive Behavior	*Total SRSS*	GPA
Angel, Julio	2310	0	0	0	2	0	0	1	3	3.1
Akins, J'Monte	2013	0	0	0	0	0	0	0	0	4.1
Backer, Brent	2031	0	1	2	1	2	2	1	9	2.3
Boxwell, Kylie	2001							0	2	2.5
Cartright, Ashley	2152							0	4	3.2
Cox, Lucille	2002							0	0	3.9
Hankins, Erin	2017							0	2	3.7
Illio, Helen	2132							0	0	2.9
Jackson, Ronald	2003	0	1	2	2	3	2	2	12	1.7
Kemp, Patrice	2009	0	0	1	0	0	0	0	1	3.3
Parker, Stephanie	2004	0	0	0	0	1	2	0	3	2.7
Reed, Kent	2010	0	0	0	0	0	0	0	0	3.6
Sterling, Michael	2022								6	2.4
Thomas, James	2018								0	3.8
Walsh, Carter	2215	0	0	0	1	0	1	0	2	3.5

Divide training set into K folds, and apply k-fold cross-validation to set hyperparameters.

Testing set is NOT used to determine hyperparameter.

4. Train model on the training set.

Student Name	Student ID	Steal	Lie, Cheat, Sneak	Behavior Problem	Peer Rejection	Low Academic Achievement	Negative Attitude	Aggressive Behavior	*Total SRSS*	GPA
Angel, Julio	2310	0	0	0	2	0	0	1	3	3.1
Akins, J'Monte	2013	0	0	0	0	0	0	0	0	4.1
Backer, Brent	2031	0	1	2	1	2	2	1	9	2.3
Boxwell, Kylie	2001	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	2	2.5
Cartright, Ashley	2152	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	4	3.2
Cox, Lucille	2002	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	0	3.9
Hankins, Erin	2017	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	2	3.7
Illio, Helen	2132	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	0	2.9
Jackson, Ronald	2003	0	1	2	2	3	2	2	12	1.7
Kemp, Patrice	2009	0	0	1	0	0	0	0	1	3.3
Parker, Stephanie	2004	0	0	0	0	1	2	0	3	2.7
Reed, Kent	2010	0	0	0	0	0	0	0	0	3.6
Sterling, Michael	2022	0	0	1	0	3	1	1	6	2.4
Thomas, James	2018	0	0	0	0	0	0	0	0	3.8
Walsh, Carter	2215	0	0	0	1	0	1	0	2	3.5

Train your model (with preset hyperparameter)

and get a training error.

5. Test your model on testing set. Report testing error. (done!)

Student Name	Student ID	Steal	Lie, Cheat, Sneak	Behavior Problem	Peer Rejection	Low Academic Achievement	Negative Attitude	Aggressive Behavior	*Total SRSS*	GPA
Angel, Julio	2310	0	0	0	2	0	0	1	3	3.1
Akins, J'Monte	2013	0	0	0	0	0	0	0	0	4.1
Backer, Brent	2031	0	1	2	1	2	2	1	9	2.3
Boxwell, Kylie	2001	0	0	0	1	1	0	0	2	2.5
Cartright, Ashley	2152	0	1	1	1	0	1	0	4	3.2
Cox, Lucille	2002	0	0	0	0	0	0	0	0	3.9
Hankins, Erin	2017	0	0	0	0	0	2	0	2	3.7
Illio, Helen	2132	0	0	0	0	0	0	0	0	2.9
Jackson, Ronald	2003	0	1	2	2	3	2	2	12	1.7
Kemp, Patrice	2009	0	0	1	0	0	0	0	1	3.3
Parker, Stephanie	2004	0	0	0	0	1	2	0	3	2.7
Reed, Kent	2010	0	0	0	0	0	0	0	0	3.6
Sterling, Michael	2022	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	6	2.4
Thomas, James	2018	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	3.8
Walsh, Carter	2215	0	0	0	1	0	1	0	2	3.5

Test your model to get testing error.

In practice, you can repeat steps 2,4,5 multiple times (e.g., 20)

- each time, randomly split training set and testing set (75%-25%)
- each time, get a testing error; finally, report averaged testing error
- no need to repeatedly set hyper-parameters

Student Name	Student ID	Steal	Lie, Cheat, Sneak	Behavior Problem	Peer Rejection	Low Academic Achievement	Negative Attitude	Aggressive Behavior	*Total SRSS*	GPA
Angel, Julio	2310	0	0	0	2	0	0	1	3	3.1
Akins, J'Monte	2013	0	0	0	0	0	0	0	0	4.1
Backer, Brent	2031	0	1	2	1	2	2	1	9	2.3
Boxwell, Kylie	2001	0	0	0	1	1	0	0	2	2.5
Cartright, Ashley	2152	0	1	1	1	0	1	0	4	3.2
Cox, Lucille	2002	0	0	0	0	0	0	0	0	3.9
Hankins, Erin	2017	0	0	0	0	0	2	0	2	3.7
Illio, Helen	2132	0	0	0	0	0	0	0	0	2.9
Jackson, Ronald	2003	0	1	2	2	3	2	2	12	1.7
Kemp, Patrice	2009	0	0	1	0	0	0	0	1	3.3
Parker, Stephanie	2004	0	0	0	0	1	2	0	3	2.7
Reed, Kent	2010	0	0	0	0	0	0	0	0	3.6
Sterling, Michael	2022	0	0	1	0	3	1	1	6	2.4
Thomas, James	2018	0	0	0	0	0	0	0	0	3.8
Walsh, Carter	2215	0	0	0	1	0	1	0	2	3.5