LING 506 - TOPICS IN COMPUTATIONAL LINGUISTICS

Introductory Machine Learning

Yan Tang

Department of Linguistics, UIUC

Week 5

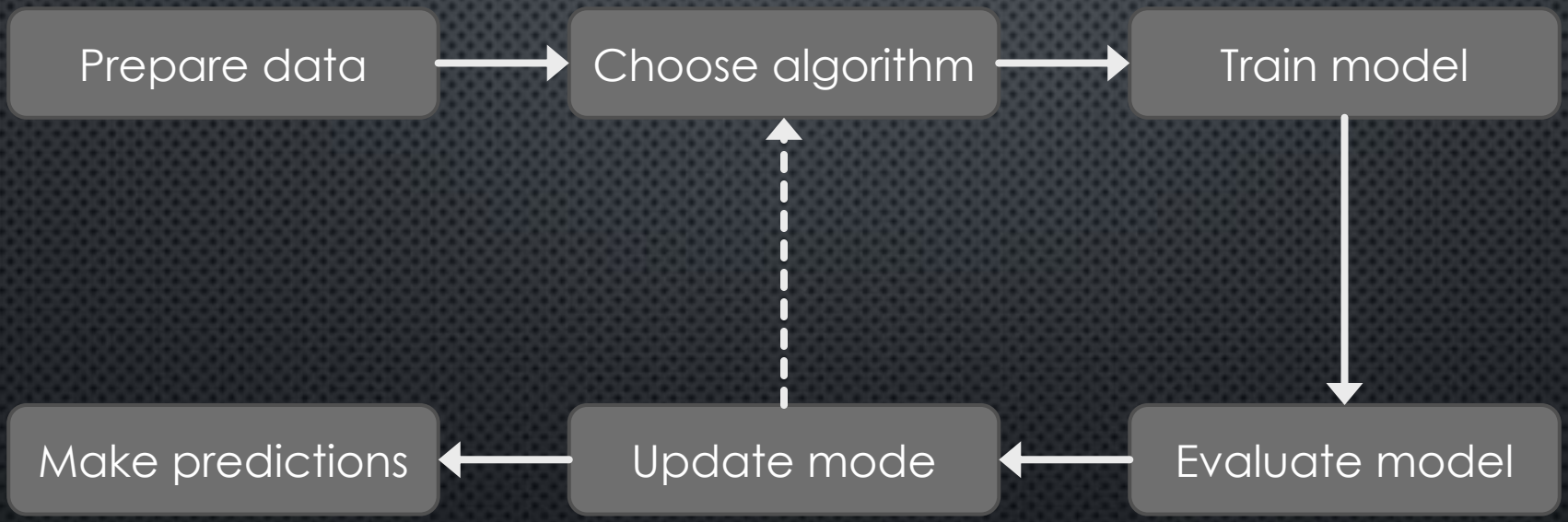
Last week…

- Data standardisation
  - Scale variables from different ranges to a comparable domain
- Parallel coordinates
  - Visualise multi-dimensional features; insights in correlations between features
- The elbow method
  - Finding k by fitting the model with a range of values

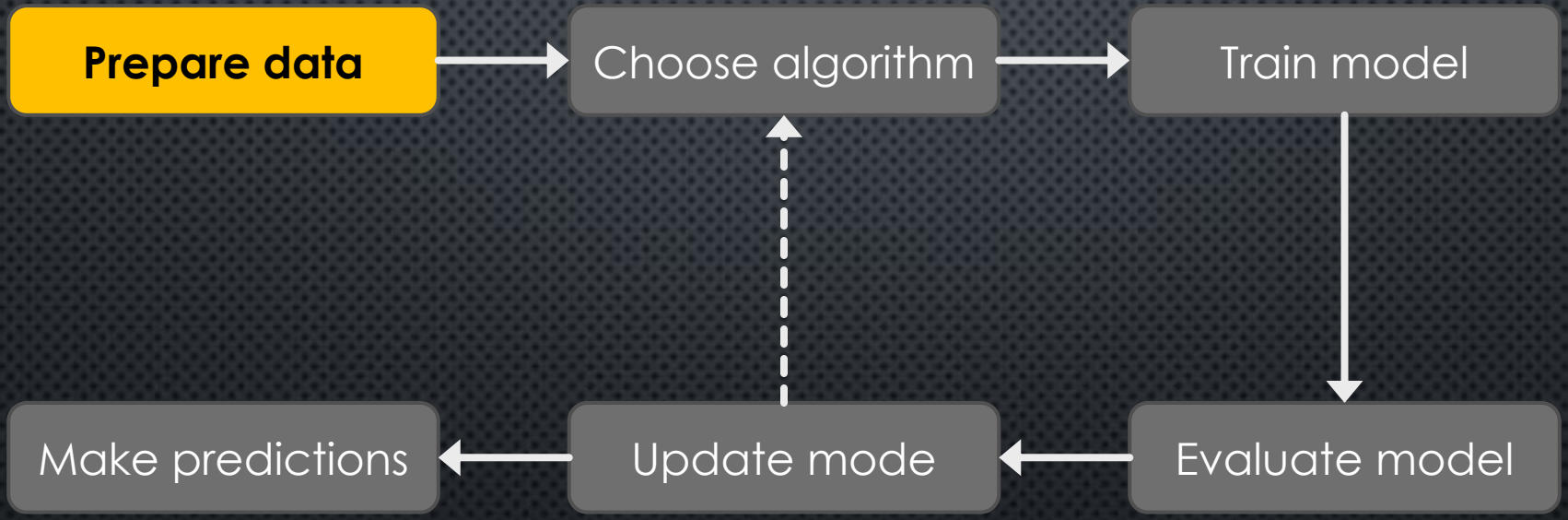
Last week…

- The Silhouette plot
  - Judge the quality of the clusters by computing Silhouette coefficient of each observation
- Hierarchical clustering
  - Represent clusters as a tree, visualised using Dengrogram
- Distance metrics
  - Euclidean (L2), Manhattan (L1) and Cosine

Building a supervised classifier or regressor



Building a supervised classifier or regressor: preparing data



Prepare data: missing data

Data cleaning: ML algorithms mostly do not work with "missing" features

	1	2	3	4
1	5.1000	3.5000	1.4000	0.2000
2	4.9000	3	1.4000	NaN
3	4.7000	3.2000	1.3000	0.2000
4	4.6000	3.1000	1.5000	0.2000
5	5	3.6000	1.4000	0.2000
6	5.4000	3.9000	1.7000	NaN
7	4.6000	NaN	1.4000	NaN
8	5	3.4000	1.5000	0.2000
9	4.4000	2.9000	1.4000	0.2000
10	4.9000	3.1000	1.5000	0.1000
11	5.4000	3.7000	1.5000	0.2000
12	4.8000	3.4000	1.6000	NaN
13	4.8000	3	1.4000	0.1000
14	4.3000	3	1.1000	0.1000

Prepare data: missing data

Data cleaning: ML algorithms mostly do not work with "missing" features

- Remove the corresponding observation/sample
- Remove the entire feature
- Fill in with artificial value(e.g. mean, median, zero, etc)

Prepare data: text and categorical data

Text and categorical data : ML algorithms prefer numbers to text:

	1 age	2 workClass	3 fnlwgt	4 education	5 education_num	6 marital_status	7 occupation	8 relationship	9 race	10 sex
1	25	Private	226802	11th	7	Never-married	Machine-op-...	Own-child	Black	Male
2	38	Private	89814	HS-grad	9	Married-civ-sp...	Farming-fishi...	Husband	White	Male
3	28	Local-gov	336951	Assoc-acdm	12	Married-civ-sp...	Protective-serv	Husband	White	Male
4	44	Private	160323	Some-college	10	Married-civ-sp...	Machine-op-...	Husband	Black	Male
5	18	<undefined>	103497	Some-college	10	Never-married	<undefined>	Own-child	White	Female
6	34	Private	198693	10th	6	Never-married	Other-service	Not-in-family	White	Male
7	29	<undefined>	227026	HS-grad	9	Never-married	<undefined>	Unmarried	Black	Male
8	63	Self-emp-n...	104626	Prof-school	15	Married-civ-sp...	Prof-specialty	Husband	White	Male
9	24	Private	369667	Some-college	10	Never-married	Other-service	Unmarried	White	Female
10	55	Private	104996	7th-8th	4	Married-civ-sp...	Craft-repair	Husband	White	Male
11	65	Private	184454	HS-grad	9	Married-civ-sp...	Machine-op-...	Husband	White	Male
12	36	Federal-gov	212465	Bachelors	13	Married-civ-sp...	Adm-clerical	Husband	White	Male
13	26	Private	82091	HS-grad	9	Never-married	Adm-clerical	Not-in-family	White	Female
14	58	<undefined>	299831	HS-grad	9	Married-civ-sp...	<undefined>	Husband	White	Male
15	48	Private	279724	HS-grad	9	Married-civ-sp...	Machine-op-...	Husband	White	Male
16	43	Private	346189	Masters	14	Married-civ-sp...	Exec-manag...	Husband	White	Male
17	20	State-gov	444554	Some-college	10	Never-married	Other-service	Own-child	White	Male
18	43	Private	128354	HS-grad	9	Married-civ-sp...	Adm-clerical	Wife	White	Female

Prepare data: text and categorical data

Text and categorical data : ML algorithms prefer numbers to text:

- Convert text to ordinal data, e.g. 0, 1, 2
- Use *one-hot* encoding
  - Create a few binary codes (0 and 1) to indicate the values in this features
  - The length of codes is equal to number of categories in the feature
  - Dummy attributes

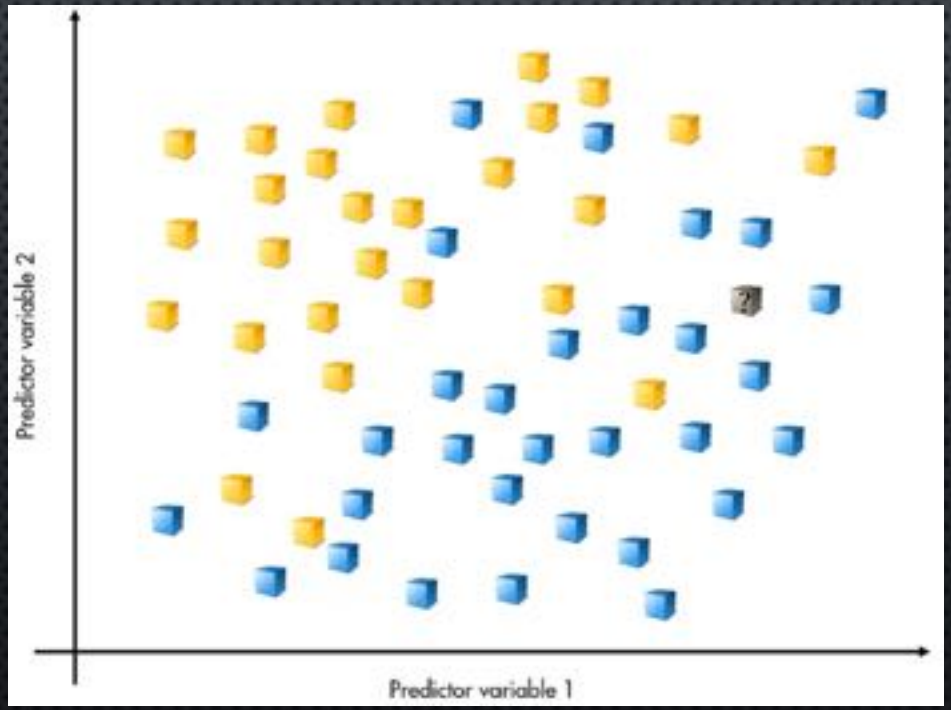
Prepare data: feature scaling

Feature scaling: ML algorithms do not perform well when numeric features fall in different scales

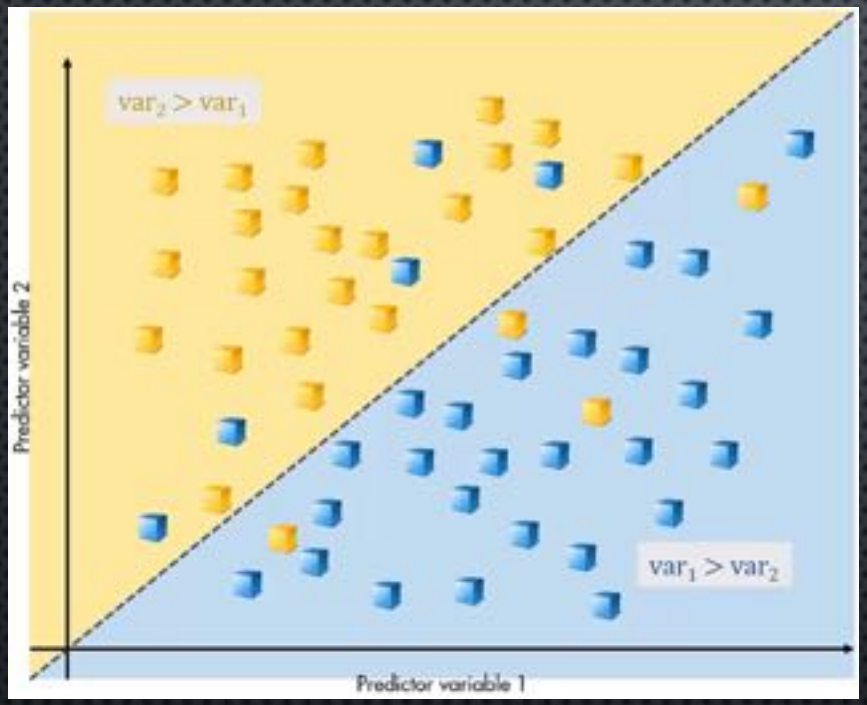
- Standardisation
- Normalisation

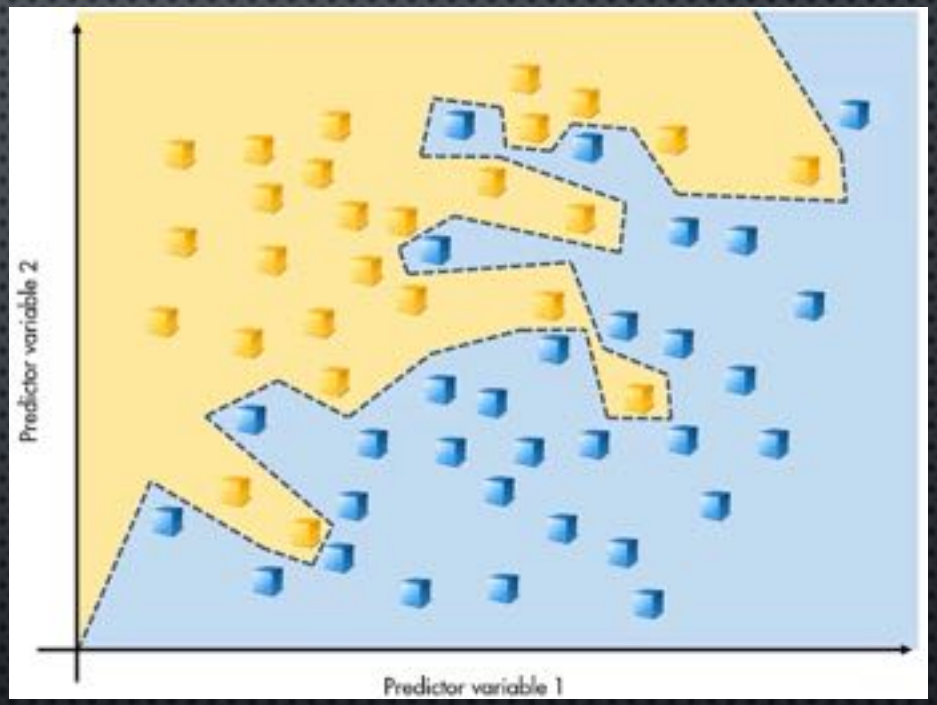
Prepare data:

How do we know if the model will accurately predict cases that were not part of the data?



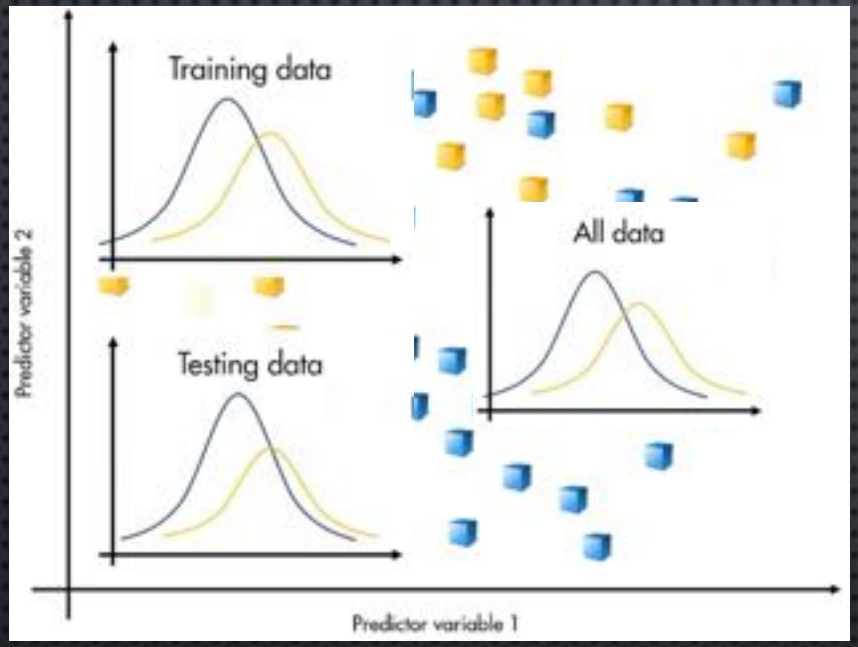
Prepare data: data partition

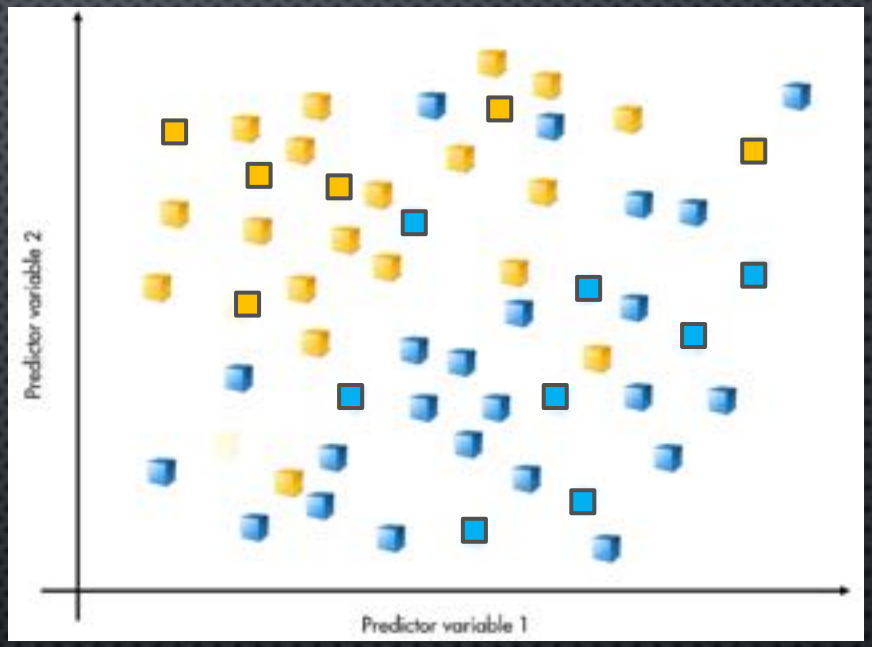




- Fit a simple model: underfitting
- Fit a complex model: overfitting

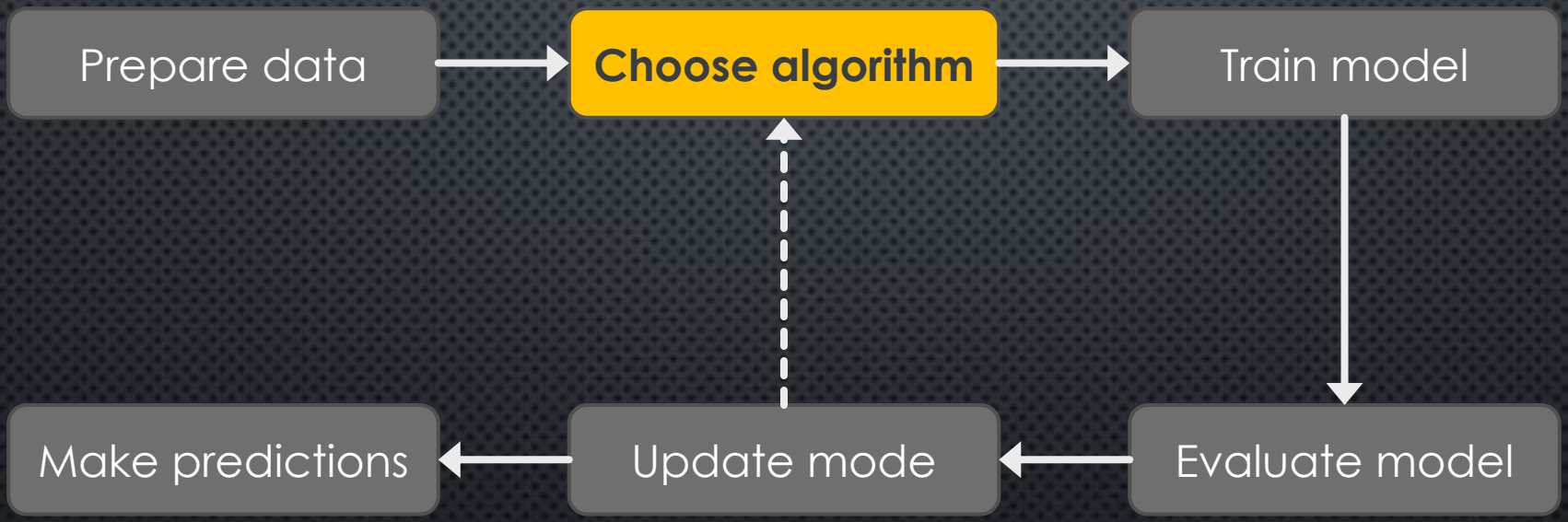
Prepare data: data partition





- Split the data into training and testing parts
- Both sets are statistically representative of the whole data set

Building a supervised classifier or regressor : choose algorithm

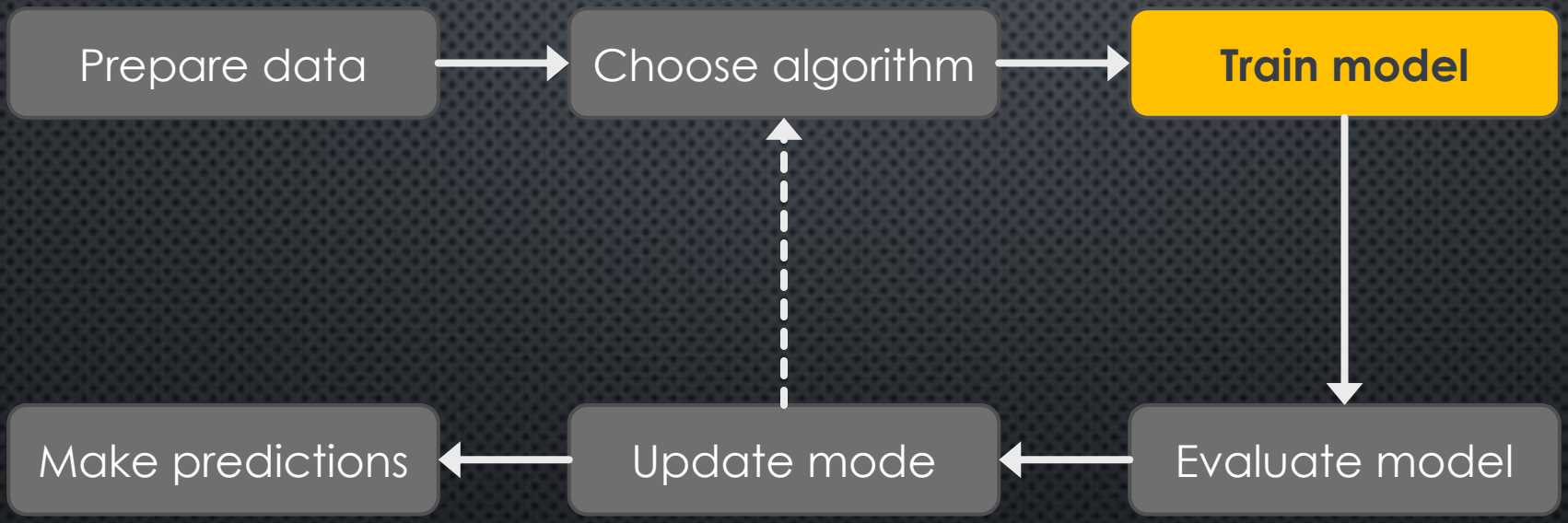


Choose algorithm: aspects to consider

After revisiting the problem and data structure:

- Complexity of model
- Accuracy of model
- Scalability of model
- Interpretability of model
- Time model takes to build, train, and test
- Speed of making predictions
- Complexity of deployment

Building a supervised classifier or regressor: train model



Train model: regression

Select a performance measure/cost function

- Regression problem
  - Root Mean Square Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - Y_i)^2}{N}}$$

y: predicted value

Y: ground truth

N: number of observations

  - Mean Square Error (MSE)

Train model: regression

- Mean Absolute Error (AME):

$$\mathrm{AME} = \frac{\sum_{i=1}^{N} |y_i - Y_i|}{N}$$

y: predicted value

Y: ground truth

N: number of observations

Train model: classification

- Classification problem
  - Categorical Cross Entropy:

$$E = \frac{\sum_{i=1}^{N} e_i}{N}$$

e : cross entropy of an observation:

$$e = -\sum_{i=1}^{K} p_i \cdot \log(p'_i)$$

K: number of classes

p: actual probability distribution of an observation

p': model-suggested probability distribution of an observation

Train model: classification

Example:

Class = {"USA", "UK", "UEA"}

$p(obv)$ = [1, 0, 0]

$p'(obv)$ = [0.75, 0.20, 0.05]

$$e = -(1 * \log(0.75) + 0 * \log(0.20) + 0 * \log(0.05)) = 0.288$$

Train model: classification

- Categorical Cross Entropy – binary problem:

$$e = \begin{cases} -p \cdot \log(p'), p = 1 \\ -(1-p) \cdot \log(1-p'), p = 0 \end{cases}$$

- Example:

  Class = {"left", "right"}

  $p(obv)$ = [0, 1]

  $p'(obv)$ = [0.75, 0.25]

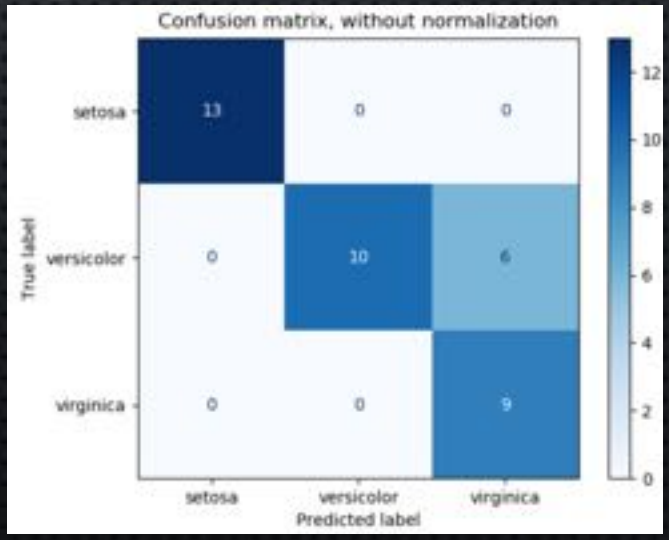
$$e = -1 * \log(1 - 0.75) = 1.386$$

Train model: classification

- Precision and recall

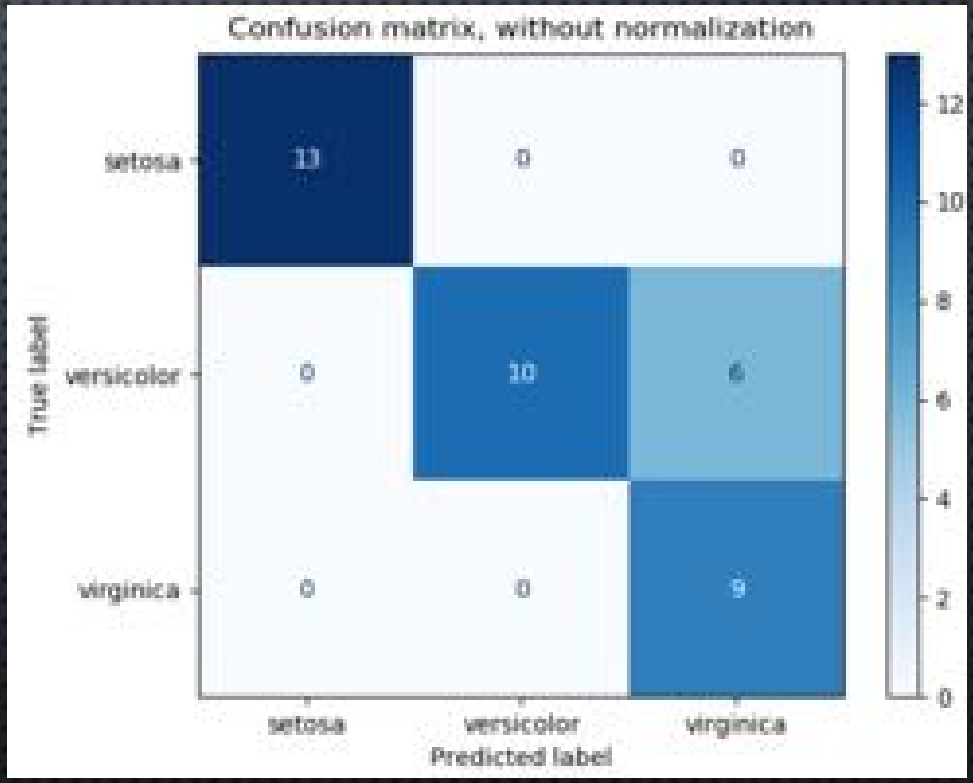
$$Precision = \frac{True\_Positive}{True\_Positive + False\_Positive}$$

$$Recall = \frac{True\_Positive}{True\_Positive + False\_Negative}$$



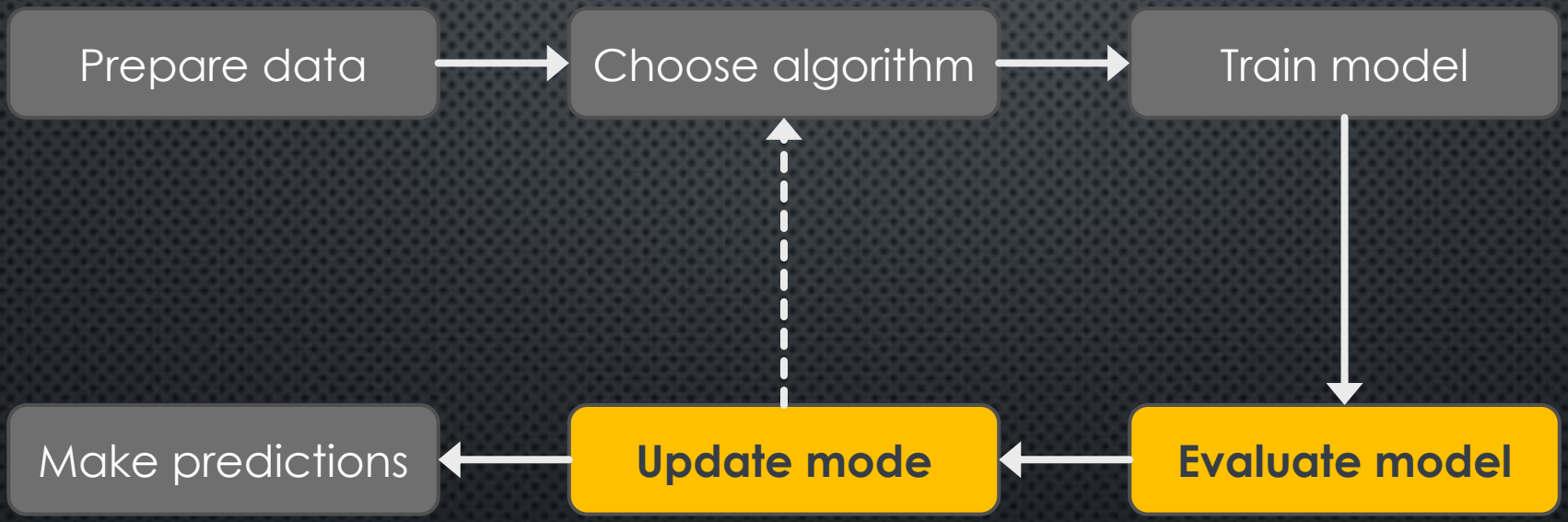
Evaluate and update model:

- Confusion matrix:
  - Classification
  - Visualise the distribution of all predicted responses and how they compare to the ground truth



https://scikit-learn.org/stable/auto_examples/model_selection/plot_confusion_matrix.html

Building a supervised classifier: evaluate model



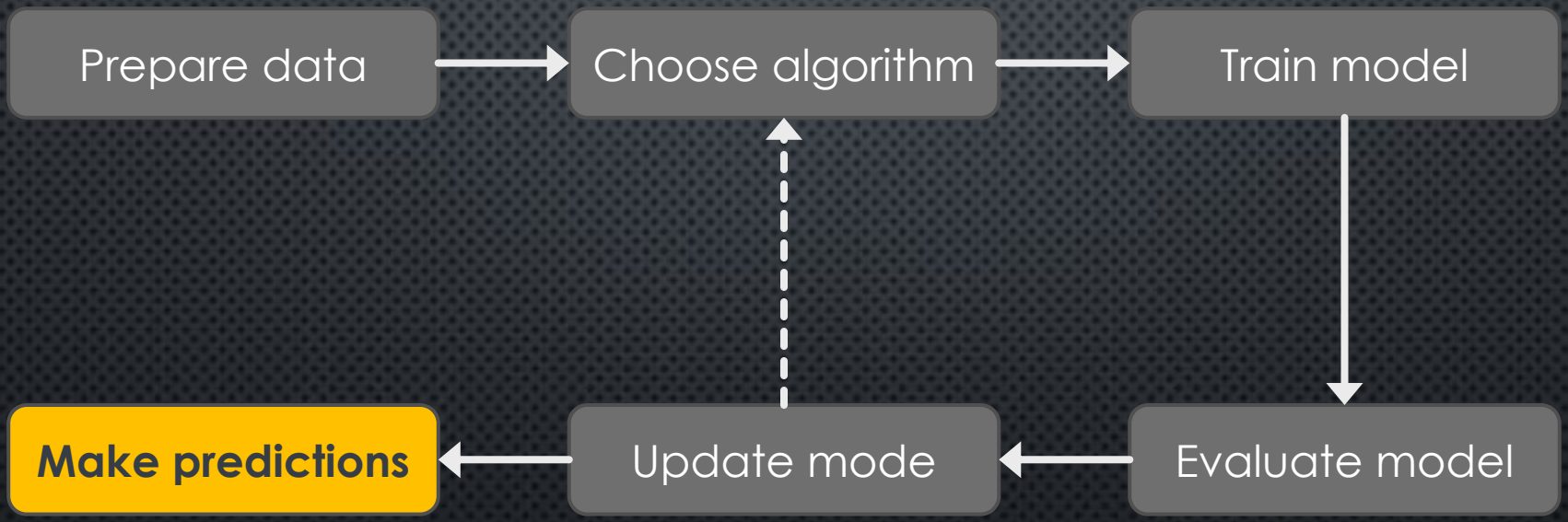
Evaluate and update model:

- Help in choosing the model that generalises well to the unknown data
- K-fold cross-validation (in three weeks!)
- Hyperparameter tuning (in three weeks!)

Evaluate and update model:

- Resubstitution:
  - Regression
  - Error between ground truth and predictions on the entire training set
  - MSE
- Hold out: validation dateset
  - Avoid the resubstitution error
  - Training data is further partitioned into two with a given ratio
  - Use equal distribution to avoid unbalance in data

Building a supervised classifier: make prediction



Review types of dataset

Training data	Validation data	Testing data