Actions

Iclicker questions and solution reorganized!!

Lecture 1

Select all of the following statements which are True (iClicker)

- (A) Predicting spam is an example of machine learning.
- (B) Predicting housing prices is not an example of machine learning.
- (C) For problems such as spelling correction, translation, face recognition, spam identification, if you are a domain expert, it's usually faster and scalable to come up with a robust set of rules manually rather than building a machine learning model.
- (D) If you are asked to write a program to find all prime numbers up to a limit, it is better to implement one of the algorithms for doing so rather than using machine learning.
- (E) Google News is likely be using machine learning to organize news.

Ans: A, D, E

Lecture 2

Select all of the following statements which are examples of supervised machine learning

- (A) Finding groups of similar properties in a real estate data set.
- (B) Predicting whether someone will have a heart attack or not on the basis of demographic, diet, and clinical measurement.
- (C) Grouping articles on different topics from different news sources (something like the Google News app).
- (D) Detecting credit card fraud based on examples of fraudulent and non-fraudulent transactions.
- (E) Given some measure of employee performance, identify the key factors which are likely to influence their performance.

ML: is mostly about predictions, not always about finding what features contribute to predictions (inference problem) - inference is mostly done with statistical approaches

Ans: B, D, E (ambiguous point can be both supervised or a stat inference problem)

Select all of the following statements which are examples of regression problems

- (A) Predicting the price of a house based on features such as number of bedrooms and the year built.
- (B) Predicting if a house will sell or not based on features like the price of the house, number of rooms, etc.
- (C) Predicting percentage grade in CPSC 330 based on past grades.
- (D) Predicting whether you should bicycle tomorrow or not based on the weather forecast.
- (E) Predicting appropriate thermostat temperature based on the wind speed and the number of people in a room.

Ans: A, C, E

B and D are classification problems

Select all of the following statements which are TRUE.

- (A) Change in features (i.e., binarizing features above) would change `DummyClassifier` predictions.
- (B) `predict` takes only `X` as argument whereas `fit` and `score` take both `X` and `y` as arguments.
- (C) For the decision tree algorithm to work, the feature values must be binary.
- (D) The prediction in a decision tree works by routing the example from the root to the leaf.

Ans: B, D (C is not true since the value cutoffs are designed during the tree learning, the features themselves don't need to be binary)

A - false. DummyClassifier doesn't look at features at all. B - true (predict we don't have values of y that's why we are predicting). C - false. Features can have values of a range %. Can work with both categorical, continuous and binary features. D - true

Lecture 3

Select all of the following statements which are TRUE.

- (A) A decision tree model with no depth is likely to perform very well on the deployment data.
- (B) Data splitting helps us assess how well our model would generalize.
- (C) Deployment data is used at the very end and only scored once.
- (D) Validation data could be used for hyperparameter optimization.
- (E) It's recommended that data be shuffled before splitting it into `train` and `test` sets.

A - false (very likely to overfit, no depth here refers to max depth set to None so it goes all the way to a complicated model) C- false (can't score deployment data we don't have access to it, we use test data)

Ans: B - true (data splitting so can do train and assess), D, E - true

With no reshuffling might have examples with all 1 or all 0

Select all of the following statements which are TRUE.

- (A) k-fold cross-validation calls fit k times.
- (B) We use cross-validation to get a more robust estimate of model performance.
- (C) If the mean train accuracy is much higher than the mean cross-validation accuracy it's likely to be a case of overfitting.
- (D) The fundamental tradeoff of ML states that as training error goes down, validation error goes up.
- (E) A decision stump on a complicated classification problem is likely to underfit.

A - true (when you do cross val on each fold we call fit (fitting a model) on the validation set)

B - true

C - true

D - false (as training error goes down, the gap between the validation error and training error goes up) not the actual validation error

As you increase model complexity, Etrain tends to go down but Evalid–Etrain tends to go up.

E - true (likely to underfit) - have high bias

Lecture 4

Select all of the following statements which are TRUE.

1. Analogy-based models find examples from the test set that are most similar to the query example we are predicting.
2. Euclidean distance will always have a non-negative value.
3. With k-NN, setting the hyperparameter k to larger values typically reduces training error.
4. Similar to decision trees, k-NNs finds a small set of good features.
5. In k-NN, with $k > 1$, the classification of the closest neighbour to the test example always contributes the most to the prediction.

1. F, T, F, F, F

1 - The models are finding examples in the training set, most similar to the test example. Not examples in the test set that are most similar to the test examples

2 - True

3 - increases training error since underfitting

4 - No - uses all the features

5 - No, it uses a majority voting unless we want to assign a weight to the distances which is not the case for k-NN

iClicker cloud join link: https://join.iclicker.com/3DP5H

Select all of the following statements which are TRUE.

1. k-NN may perform poorly in high-dimensional space (say, $d > 1000$).
2. In SVM RBF, removing a non-support vector would not change the decision boundary.
3. In sklearn's SVC classifier, large values of gamma tend to result in higher training score but probably lower validation score.
4. If we increase both gamma and C, we can't be certain if the model becomes more complex or less complex.

T, T, T, F (because both increase in the same direction)

Lecture 5

1. `StandardScaler` ensures a fixed range (i.e., minimum and maximum values) for the features.
2. `StandardScaler` calculates mean and standard deviation for each feature separately.
3. In general, it's a good idea to apply scaling on numeric features before training k-NN or SVM RBF models.
4. The transformed feature values might be hard to interpret for humans.
5. After applying `SimpleImputer` The transformed data has a different shape than the original data.

B, C, D (True)

A, E - False returns a numpy array but same shape since imputing just fills in missing values

Select all of the following statements which are TRUE.

1. You can have scaling of numeric features, one-hot encoding of categorical features, and `scikit-learn` estimator within a single pipeline.
2. Once you have a `scikit-learn` pipeline object you can call `fit`, `predict`, and `score` on it.
3. You can carry out data splitting within `scikit-learn` pipeline.
4. We have to be careful of the order we put each transformation and model in a pipeline.
5. Pipelines will `fit` and `transform` on the training fold and only `transform` on the validation fold during cross-validation.

1 - False (will learn this more later - no we cant do this in a single pipeline)

2 - True (if it doesn't have estimator as the last object can't call predict and score on it, it's only a pipeline with transformation)

3 - False (it's not a transformation in data, before we start training our model we have to split the data)

4 - True (very important that we put the order of transformation in the right order, applied sequentially on the data)

5 - True (this is the purpose of our pipeline object)

Lecture 6

6.1

iClicker cloud join link: https://join.iclicker.com/3DP5H

Select all of the following statements which are TRUE.

1. You could carry out cross-validation by passing a `ColumnTransformer` object to `cross_validate`.
2. After applying column transformer, the order of the columns in the transformed data has to be the same as the order of the columns in the original data.
3. After applying a column transformer, the transformed data is always going to be of different shape than the original data.
4. When you call `fit_transform` on a `ColumnTransformer` object, you get a numpy ndarray.

F - we don't have an estimator in column transformer so cant do fit, predict

F - any order we like (depends in the order we create our column transformer object)

F - not always of different shape (if only numeric feat and pass through)

True - we get a numpy array

Select all of the following statements which are TRUE.

- (A) `handle_unknown="ignore"` would treat all unknown categories equally.
- (B) As you increase the value for `max_features` hyperparameter of `CountVectorizer` the training score is likely to go up.
- (C) Suppose you are encoding text data using `CountVectorizer`. If you encounter a word in the validation or the test split that's not available in the training data, we'll get an error.
- (D) In the code below, inside `cross_validate`, each fold might have slightly different number of features (columns) in the fold.

```
pipe = (CountVectorizer(), SVC())
cross_validate(pipe, X_train, y_train)
```

True - unknown category so will make it 000 for all extracted feature (not add a new column)

True - it gets all the vocabulary

False - After filtering our vocab is only the words we have extracted. Now in our new review - it ignores all the words not in our vocab. if a new word shows up, I found unique words represent my test docment in terms of my vocab.

True - it is possible since Count Vectorizer extracts a different number of vocabs each time

Lecture 7

Select all of the following statements which are TRUE.

- (A) Increasing the hyperparameter `alpha` of `Ridge` is likely to decrease model complexity.
- (B) `Ridge` can be used with datasets that have multiple features.
- (C) With Ridge, we learn one coefficient per training example.
- (D) If you train a linear regression model on a 2-dimensional problem (2 features), the model will learn 3 parameters: one for each feature and one for the bias term.

A - True (smaller values of alpha, it likely to overfit. larger values of alpha very low training score so under fitted model)

Ridge - selects all features, just makes coefficients very small for some features

B - True

C - False - dont learn coeff with a example its with each feature

D - True

Select all of the following statements which are TRUE.

- (A) Increasing logistic regression's `C` hyperparameter increases model complexity.
- (B) The raw output score can be used to calculate the probability score for a given prediction.
- (C) For linear classifier trained on d features, the decision boundary is a $d-1$-dimensional hyperparlane.
- (D) A linear model is likely to be uncertain about the data points close to the decision boundary.

A - True

Increasing complexity - the coefficients have bigger value

Less complex values - coefficients smaller

B - True - raw scores after calculating each feature value with coeff + bias. apply sigmoid to it to get probability scores

C - true. 2 features - decision boundary is going to be a line, 3 features a plane

D - True - close to decision boundary uncertain about prediction

Lecture 8

iClicker cloud join link: https://join.iclicker.com/3DP5H

Select all of the following statements which are TRUE.

- (A) If you get best results at the edges of your parameter grid, it might be a good idea to adjust the range of values in your parameter grid.
- (B) Grid search is guaranteed to find best hyperparameters values.
- (C) It is possible to get different hyperparameters in different runs of `RandomizedSearchCV`.

T (explore the space more), F (We are giving it a range on our own, best might not be here), T (yes we can)

Questions for class discussion (hyperparameter optimization)

- Suppose you have 10 hyperparameters, each with 4 possible values. If you run `GridSearchCV` with this parameter grid, how many cross-validation experiments will be carried out?
- Suppose you have 10 hyperparameters and each takes 4 values. If you run `RandomizedSearchCV` with this parameter grid with `n_iter=20`, how many cross-validation experiments will be carried out?

Answers -

- 4^10 different combinations

- 20 (randomly selects 20 combinations - so better to use distributions)

Questions for whether we should trust a model:

1. Probably not - no all as train set so overfitting
2. Probably yes - most likely yes
3. Probably not - too small - overfitting of validation set

Lecture 9

Select all of the following statements which are TRUE.

- (A) In medical diagnosis, false positives are likely to be more damaging than false negatives (assume "positive" means the person has a disease, "negative" means they don't).
- (B) In spam classification, false positives are more damaging than false negatives (assume "positive" means the email is spam, "negative" means it's not).
- (C) If method A gets a higher accuracy than method B, that means its precision is also higher.
- (D) If method A gets a higher accuracy than method B, that means its recall is also higher.

A - less damaging (False)

B - it is True - in case of emails it means an imp email has been marked as spam - very risky might lose imp emails

if false positive is more important - metric we should care about: minimize false positive so have higher precision. If false neg more important - recall is more imp

C - False (not necessary)

D - False (not necessarily True)

Lecture 10

Select all of the following statements which are TRUE.

- (A) Price per square foot would be a good feature to add in our `X`.
- (B) The `alpha` hyperparameter of `Ridge` has similar interpretation of `C` hyperparameter of `LogisticRegression`; higher `alpha` means more complex model.
- (C) In regression, one should use MAPE instead of MSE when relative (percent) error matters more than absolute error.
- (D) A lower RMSE value indicates a better model.
- (E) We can use still use precision and recall for regression problems but now we have other metrics we can use as well.

A - no point since we can just have directly our target (False)

B - false - opp interpretation

C - True

D - True

E - false - cannot use precision recall

Select all of the following statements which are TRUE.

- (A) Price per square foot would be a good feature to add in our `X`.
- (B) The `alpha` hyperparameter of `Ridge` has similar interpretation of `C` hyperparameter of `LogisticRegression`; higher `alpha` means more complex model.
- (C) In `Ridge`, smaller alpha means bigger coefficients whereas bigger alpha means smaller coefficients.

True statements:

- C

Select all of the following statements which are TRUE.

- (A) We can use still use precision and recall for regression problems but now we have other metrics we can use as well.
- (B) In `sklearn` for regression problems, using `r2_score()` and `.score()` (with default values) will produce the same results.
- (C) RMSE is always going to be non-negative.
- (D) MSE does not directly provide the information about whether the model is underpredicting or overpredicting.
- (E) We can pass multiple scoring metrics to `GridSearchCV` or `RandomizedSearchCV` for regression as well as classification problems.

True statements:

- B

- C

- D

- E

Lecture 11

Select all of the following statements which are TRUE.

- (A) Every tree in a random forest uses a different bootstrap sample of the training set.
- (B) To train a tree in a random forest, we first randomly select a subset of features. The tree is then restricted to only using those features.
- (C) A reasonable implementation of `predict_proba` for random forests would be for each tree to "vote" and then normalize these vote counts into probabilities.
- (D) Increasing the hyperparameter max_features (the number of features to consider for a split) makes the model more complex and moves the fundamental tradeoff toward lower training error.
- (E) A random forest with only one tree is likely to get a higher training error than a decision tree of the same depth.

A - True

B - False not at tree level, at node level

C - True D - True (increasing complexity so training error goes down)

E - True - having a subset of my training set and subset of features so less overfit so higher training error

Select the most accurate option below.

- (A) Every tree in a random forest uses a different bootstrap sample of the training set.
- (B) To train a tree in a random forest, we first randomly select a subset of features. The tree is then restricted to only using those features.
- (C) The `n_estimators` hyperparameter of random forests should be tuned to get a better performance on the validation or test data.
- (D) In random forests we build trees in a sequential fashion, where the current tree is dependent upon the previous tree.
- (E) Let classifiers A, B, and C have training errors of 10%, 20%, and 30%, respectively. Then, the best possible training error from averaging A, B and C is 10%.

True statements:

- A

Lecture 13

iClicker cloud join link: https://join.iclicker.com/3DP5H

Select all of the following statements which are TRUE.

- (A) Simple association-based feature selection approaches do not take into account the interaction between features.
- (B) You can carry out feature selection using linear models by pruning the features which have very small weights (i.e., coefficients less than a threshold).
- (C) Forward search is guaranteed to find the best feature set.
- (D) The order of features removed given by `rfe.ranking_` is the same as the order of original feature importances given by the model.

True - no feature target, just correlation no modelling only linear association

True - Model feature based selection

False - no its not guaranteed

False - When a feature is removed, the feature importance for other features can go up or down. So for rfe.ranking_ the features are removed iteratively and can be different from the original feature importance since that is determined on the whole set

Lecture 14

iClicker cloud join link: https://join.iclicker.com/SNBF

- (A) K-Means is sensitive to initialization and the solution may change depending upon the initialization.
- (B) K-means terminates when the number of clusters does not increase between iterations.
- (C) K-means terminates when the centroid locations do not change between iterations.
- (D) K-Means is guaranteed to find the optimal solution.

True statements: A, C

Lecture 15

(iClicker) Exercise 15.1

iClicker cloud join link: https://join.iclicker.com/3DP5H

Select all of the following statements which are TRUE.

- (A) K-Means may converge to different solutions depending upon the initialization.
- (B) K-means terminates when the number of clusters does not increase between iterations.
- (C) K in K-Means should always be $\leq$ # of features.
- (D) In K-Means, it makes sense to have $K \leq$ # of examples.
- (E) In K-Means, in some iterations some points may be left unassigned.

True

False - number of clusters dont change between iteration, before we start our model training we decide k

False (no relationship with features)

True - cant definitely have more than,should be strictly <

False - each point needs to be assigned to a cluster, if we have outliers in data our cluster centres will be affected

Select all of the following statements which are TRUE.

- (A) The preprocessing methods such as `StandardScaler` are unsupervised methods.
- (B) K-means terminates when the centroid locations do not change between iterations.
- (C) If you train K-Means with `n_clusters` = the number of examples, the inertia value will be 0.
- (D) Unlike the Elbow method, the Silhouette method is not dependent on the notion of cluster centers.

True - not make use of *y* in supervised *model either*

True

True, the intracluster distance is 0

True - Elbow method - uses cluster centre, silhouette only uses cluster distance

iClicker cloud join link: https://join.iclicker.com/SNBF

- (A) If you train K-Means with `n_clusters` = the number of examples, the inertia value will be 0.
- (B) The elbow plot shows the tradeoff between within cluster distance and the number of clusters.
- (C) Unlike the Elbow method, the Silhouette method is not dependent on the notion of cluster centers.
- (D) The elbow plot is not a reliable method to obtain the optimal number of clusters in all cases.
- (E) The Silhouette scores ranges between -1 and 1 where higher scores indicates better cluster assignments.

True statements: A, B, C, D, E

Lecture 16

❓❓ Questions for you

(iClicker) Exercise 16.1

iClicker cloud join link: https://join.iclicker.com/3DP5H

Select all of the following statements which are TRUE.

- (A) With tiny epsilon (`eps` in `sklearn`) and min samples=1 (`min_samples=1` in `sklearn`) we are likely to end up with each point in its own cluster.
- (B) With a smaller value of `eps` and larger number for `min_samples` we are likely to end up with a one big cluster.
- (C) K-Means is more susceptible to outliers compared to DBSCAN.
- (D) In DBSCAN to be part of a cluster, each point must have at least `min_samples` neighbours in a given radius (including itself).
- (E) In DBSCAN, it is generally a good idea to run DBSCAN with a large number of different random orderings of training examples.

A - True (if small eps, very conservative about who else is my neighbour, min samples each point so each is on their own cluster)

small eps but big min samples - most noise points

B - False

C - True (it is, outlier affects cluster centres, DBSCAN can identify it as noise point)

D - False (only core points need min samples neighbours within eps, border points belong to cluster but not have min samples - start with a random point, determine if it has min sample within epsilon and spread the colour, repeat for the neighbours - some neighbours might not have min sample within radius then they become border points)

E - False - random starting point will not change the end result of my algo, kmeans will get changed - some border points

(iClicker) Exercise 16.2

iClicker cloud join link: https://join.iclicker.com/3DP5H

Select all of the following statements which are TRUE.

- (A) In hierarchical clustering we do not have to worry about initialization.
- (B) Hierarchical clustering can only be applied to smaller datasets because dendrograms are hard to visualize for large datasets.
- (C) In all the three clustering methods we saw (K-Means, DBSCAN, hierarchical clustering), there is a way to decide the granularity of clustering (i.e., how many clusters to pick).
- (D) To get robust clustering we can naively ensemble cluster labels (e.g., pick the most popular label) produced by different clustering methods.
- (E) If you have a high Silhouette score and very clean and robust clusters, it means that the algorithm has captured the semantic meaning in the data of our interest.

True statements: A, C

Lecture 19

iClicker Exercise 19.1

iClicker cloud join link: https://join.iclicker.com/3DP5H

Select all of the following statements which are TRUE.

- (A) It's possible to use word representations for text classification instead of bag-of-words representation.
- (B) The topic model approach we used in the last lecture, Latent Dirichlet Allocation (LDA), is an unsupervised approach.
- (C) In an LDA topic model, the same word can be associated with two different topics with high probability.
- (D) In an LDA topic model, a document is a mixture of multiple topics.
- (E) If I train a topic model on a large collection of news articles with K = 10, I would get 10 topic labels (e.g., sports, culture, politics, finance) as output.

A - (word2vec embedding instead of word representation)

True not direct, but indirect

cannot directly use them but can use to get a document representation from word representation

B - T

C - T

D - T

E - F it gives 10 topics, but not actual labels

Select all of the following statements which are TRUE.

- (A) One-vs.-one strategy uses all the available data when training each binary classifier.
- (B) For a 100-class classification problem, one-vs.-rest multi-class strategy will create 100 binary classifiers.

A - No - only consider data for class 1, class 2

B - True, creates 100

Lecture 20

Select all of the following statements which are TRUE.

- (A) We need to be careful when splitting the data when working with time series data.
- (B) Cross-validation in time series can be randomly applied like in other machine learning tasks.
- (C) In time series forecasting, the future value of a series can only be predicted based on its past values and cannot incorporate other variables.
- (D) When we used `RandomForestRegressor` model on the POSIX time feature, it predicted a straight line on the test data because tree-based models are inherently unable to extrapolate (i.e., make predictions outside the range of the training data).

True statements:

- A, D

Lecture 21

Select all of the following statements which are TRUE.

- (A) Right censoring occurs when the endpoint of event has not been observed for all study subjects by the end of the study period.
- (B) Right censoring implies that the data is missing completely at random.
- (C) In the presence of right-censored data, binary classification models can be applied directly without any modifications or special considerations.
- (D) If we apply the `Ridge` regression model to predict tenure in right censored data, we are likely to underestimate it because the tenure observed in our data is shorter than what it would be in reality.

True statements:

- A, D

followup discussions, *for lingering questions and comments*

@838_f1

Anonymous Gear 4 weeks ago

I think this is missing questions 14.1 and 14.3 from lecture 14. Any chance this could be updated?

helpful! | 0

Anonymous Poet 4 weeks ago

The questions in 14.1 and 14.3 appear throughout the Lecture 15 questions in this post

helpful! | 0

Reply to this followup discussion

@838_f2

Anonymous Helix 4 weeks ago

Anyone have the solution to these? Thanks!

Questions for you

iClicker cloud join link: https://join.iclicker.com/SNBF

- (A) Colloborative filtering is unsupervised learning
- (B) For an NxM utility matrix collobrative filtering trains N classifiers
- (C) Content based filtering can assign ratings to new users for existing items
- (D) Content based filtering can assign ratings to new items for existing users
- (E) Each user in content based filtering will have a different number of features to train on

helpful! | 1

Anonymous Beaker 4 weeks ago

I think the answers are A and D but I could be wrong

~ An instructor (Varada Kolhatkar) thinks this is a good comment ~

helpful! | 1

Anonymous Comp 3 weeks ago

Shouldn't E be true as well? Because for content- based filtering, we have additional features, right? If even one of these features is categorical, using one-hot encoder, we might get different number of features overall. In the example we had in the lecture, however, we were working with already one-hot encoded features, so I'm not really sure.

helpful! | 0

Anonymous Beaker 3 weeks ago

in content-based filtering, the number of features used to represent items is consistent across all users. The system creates a user profile based on the features of the items they have interacted with or liked, but the feature space itself remains the same for everyone. It's the values or weights assigned to these features that vary from user to user, reflecting individual preferences.

helpful! | 0

Reply to this followup discussion

@838_f3

KTR 4 weeks ago

Thank you, this is very helpful.

helpful! | 0

Reply to this followup discussion

Start a new followup discussion

Compose a new followup discussion