Quick Check

11/2/2016

How many principal components should you use after doing PCA?

How many principal components should you use after doing PCA?

However many you need to explain most of the variance in your data

What's the difference between regression and classification?

What's the difference between regression and classification?

Regression predicts a real value, classification predicts a class

What's a silhouette score?

What's a silhouette score?

A measure of how closely related a point is to members of its cluster vs those of other clusters

Why would you use regularization? Explain L1 and L2 regularization.

Why would you use regularization? Explain L1 and L2 regularization.

- **To penalize model complexity and reduce overfitting**
- **L1 :** Penalizing a model's score by including a function of the sum of absolute values of its parameter coefficients
- **L2 :** Penalizing a model's score by including a function of the sum of squared values of its parameter coefficients

Consider the following schema:

```
STUDENTS(student_code, first_name, last_name, email,
         phone_no, date_of_birth, honours_subject, percentage_of_marks);
```

Write a query that would display names of all the students whose honours subject is English, or Spanish and percentage of marks more than 80.

Consider the following schema:

```
STUDENTS(student_code, first_name, last_name, email,
         phone_no, date_of_birth, honours_subject, percentage_of_marks);
```

Write a query that would display names of all the students whose honours subject is English, or Spanish and percentage of marks more than 80.

```
SELECT first_name, last_name FROM students

WHERE (honours_subject = “English” OR honours_subject = “Spanish”)

 AND percentage_of_marks > 80;
```

How does a kNN algorithm work?

How does a kNN algorithm work?

It classifies a new observation based on a vote of its k-nearest labeled neighbors in its feature space.

Is it better to have too many false positives, or too many false negatives? Explain.

Is it better to have too many false positives, or too many false negatives? Explain.

Depends on the context!

Ex 1:

- It is better that ten guilty persons escape than that one innocent suffer?
- Is reducing the chance that a patient with cancer is mistakenly diagnosed as healthy more important?

Is it better to have too many false positives, or too many false negatives? Explain.

Depends on the context!

Ex 2:

Should startups prefer false positives or false negatives in hiring?

False positive = hiring incompetent candidates, but also hiring some hidden gems.

False negative = rejecting highly competent candidates, but avoiding people who are subtly incompetent.

Which side should startups optimize?

The "hire fast and fire fast" mantra would indicate leaning towards hiring false positives.

What is the difference between Type I error and Type II error?

What is the difference between Type I error and Type II error?

- **Type I error:** Rejecting the null hypothesis when it is in fact true
- **Type II error:** Failing to reject the null hypothesis when it is in fact false

What is a confusion matrix?

What is a confusion matrix?

A table that is used to describe the performance of a classification model on a set of test data for which the true values are known.

What's a significance level / a-value?

What's a significance level / a-value?

The probability of a Type I error

What are pros and cons of decision tree models?

What are pros and cons of decision tree models?

- **Pros:** white box; little data prep needed; robust and non-parametric; trains relatively quickly; once trained, prediction is very fast.
- **Cons:** overly complex trees overfit easily; greedy (may not find the best tree); instability (decision tree changes when the dataset is altered)

What's a dummy variable?

What's a dummy variable?

Recoding categorical variables that have more than two categories into a series of binary variables.

Explain true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

- **true positives (TP):** These are cases in which the model predicted yes correctly (e.g., they have the disease)
- **true negatives (TN):** The model predicted no correctly
- **false positives (FP):** The model predicted yes incorrectly (aka: Type I error)
- **false negatives (FN):** The model predicted no incorrectly (aka: Type II error)

Explain accuracy and misclassification rate.

- **Accuracy:** how often is the classifier correct?
  - (TP+TN)/total = (100+50)/165 = 0.91
- **Misclassification Rate:** how often is it wrong?
  - (FP+FN)/total = (10+5)/165 = 0.09
  - equivalent to 1 minus Accuracy
  - also known as "Error Rate"

Explain what precision and recall are. How do they relate to the ROC curve?