Final Prep

CPSC 330

Machine Learning Fundamentals

Which of the following **corresponds the most** to overfitting?

A) Breaking the golden rule
B) High bias
C) High variance
D) Low complexity

Machine Learning Fundamentals

Which of the following **corresponds the most** to overfitting?

A) Breaking the golden rule
B) High bias
C) High variance
D) Low complexity

Supervised Learning Models

Which model is most appropriate for a **regression** problem?

A) LogisticRegression
B) DecisionTreeClassifier
C) KNeighborsClassifier
D) Ridge

Supervised Learning Models

Which model is most appropriate for a **regression** problem?

A) LogisticRegression
B) DecisionTreeClassifier
C) KNeighborsClassifier
D) Ridge

Preprocessing

Which preprocessing step has the potential to change the shape of the dataset?

A) OneHotEncoder
B) StandardScaler
C) SimpleImputer
D) None of the above

Preprocessing

Which preprocessing step has the potential to change the shape of the dataset?

A) OneHotEncoder
B) StandardScaler
C) SimpleImputer
D) None of the above

Preprocessing

What should you do if you discover that some of your examples are missing target values?

A) Use imputation on the target column
B) Drop the target column
C) Pick a different column to predict
D) Drop rows that are missing target values

Preprocessing

What should you do if you discover that some of your examples are missing target values?

A) Use imputation on the target column
B) Drop the target column
C) Pick a different column to predict
D) Drop rows that are missing target values

Linear Models

Which statement about linear models is **true**?

A) With Ridge, we learn one coefficient per training example.
B) For a given example, predict_proba lets us see how likely each feature is to affect the final prediction.
C) Increasing a LogisticRegression model's C hyperparameter increases model complexity.
D) For a model trained on unscaled data, if Feature A's coefficient is 1.25 and Feature B's coefficient is 0.56, then Feature A must have a bigger impact on the final prediction.

Linear Models

Which statement about linear models is **true**?

A) With Ridge, we learn one coefficient per training example.
B) For a given example, predict_proba lets us see how likely each feature is to affect the final prediction.
C) Increasing a LogisticRegression model's C hyperparameter increases model complexity.
D) For a model trained on unscaled data, if Feature A's coefficient is 1.25 and Feature B's coefficient is 0.56, then Feature A must have a bigger impact on the final prediction.

Multiple Choice

If you fill out a test with 10 multiple choice (a, b, c, d) questions completely at random, your expected grade is 25%. If you fill out two tests, but only submit the better one, your expected grade becomes 33%. The more midterms you fill out, only submitting the best one each time, the more your expected grade increases. This is an example of...

A) Training a regression model
B) The fundamental trade-off
C) Optimization bias
D) The golden rule

Credit to CPSC 340 for this example :)

Multiple Choice

If you fill out a test with 10 multiple choice (a, b, c, d) questions completely at random, your expected grade is 25%. If you fill out two tests, but only submit the better one, your expected grade becomes 33%. The more midterms you fill out, only submitting the best one each time, the more your expected grade increases. This is an example of...

A) Training a regression model
B) The fundamental trade-off
C) Optimization bias
D) The golden rule

Credit to CPSC 340 for this example :)

Classification Metrics

Which metric can be used to measure a binary classification model's ability to distinguish between classes across all possible thresholds?

A. Precision
B. Recall
C. AUC-ROC
D. F1 Score

Classification Metrics

Which metric can be used to measure a binary classification model's ability to distinguish between classes across all possible thresholds?

A. Precision
B. Recall
C. AUC-ROC
D. F1 Score

Regression Metrics

In regression, which pair of metrics is most sensitive to large outliers, and which is more robust to them?

(MAE: Mean Absolute Error, RMSE: Root Mean Squared Error, MAPE: Mean Absolute Percentage Error)

A. MAE is sensitive; R^2 is robust
B. RMSE is sensitive; MAE is robust
C. R^2 is sensitive; RMSE is robust
D. MAPE is sensitive; RMSE is robust

Regression Metrics

In regression, which pair of metrics is most sensitive to large outliers, and which is more robust to them?

(MAE: Mean Absolute Error, RMSE: Root Mean Squared Error, MAPE: Mean Absolute Percentage Error)

A. MAE is sensitive; R^2 is robust
B. RMSE is sensitive; MAE is robust
C. R^2 is sensitive; RMSE is robust
D. MAPE is sensitive; RMSE is robust

Ensembles

Which of the following correctly describes the sources of randomness in a Random Forest?

A. Bootstrapped sampling of training data only
B. Random selection of features at each split only
C. Both A and B
D. Random initialization of leaf predictions

Ensembles

Which of the following correctly describes the sources of randomness in a Random Forest?

A. Bootstrapped sampling of training data only
B. Random selection of features at each split only
C. Both A and B
D. Random initialization of leaf predictions

Feature Importance

Which statement about SHAP values is TRUE?

A. SHAP values show global importance but cannot explain individual predictions
B. SHAP values compute importance by removing each feature entirely.
C. SHAP values provide additive, local explanations for individual predictions.
D. SHAP values are available only for linear models.

Feature Importance

Which statement about SHAP values is TRUE?

A. SHAP values show global importance but cannot explain individual predictions
B. SHAP values compute importance by removing each feature entirely.
C. SHAP values provide additive, local explanations for individual predictions.
D. SHAP values are available only for linear models.

Feature Importance

You use permutation importance on a dataset where many features are highly correlated. What pattern should you expect?

A. Each correlated feature gets extremely high importance.
B. One feature in each correlated group gets most of the importance; others get little.
C. All correlated features are ranked equally.
D. Permutation importance fails and returns NaN values

Feature Importance

You use permutation importance on a dataset where many features are highly correlated. What pattern should you expect?

A. Each correlated feature gets extremely high importance.
B. One feature in each correlated group gets most of the importance; others get little.
C. All correlated features are ranked equally.
D. Permutation importance fails and returns NaN values

Feature Engineering

You are engineering features for a tree-based ensemble. Which transformation is generally *less* necessary for tree models compared to linear models, and why?

A. One-hot encoding of categorical variables
B. Standardizing continuous features
C. Ordinal encoding of categorical features
D. Creating interaction terms (e.g., X_1^2, X_1X_2)

Feature Engineering

You are engineering features for a tree-based ensemble. Which transformation is generally *less* necessary for tree models compared to linear models, and why?

A. One-hot encoding of categorical variables
B. Standardizing continuous features
C. Ordinal encoding of categorical features
D. Creating interaction terms (e.g., X_1^2, X_1X_2)

Clustering

A biologist wants to group gene expression samples but has no idea how many clusters there might be. Which clustering method most easily allows her to try out different levels of granularity?

A. K-Means
B. DBSCAN
C. Hierarchical clustering
D. None of the above

Clustering

A biologist wants to group gene expression samples but has no idea how many clusters there might be. Which clustering method most easily allows her to try out different levels of granularity?

A. K-Means
B. DBSCAN
C. Hierarchical clustering
D. None of the above

Recommender Systems

Which of the following is NOT true about content-based filtering?

A. It can be used to predict a customer's expected rating of a new item (one without any existing ratings from other users)
B. It requires more investment in feature acquisition and engineering
C. It is a supervised learning approach
D. It is not transparent – i.e., it is not possible to say for sure why an item was recommended to a user

Recommender Systems

Which of the following is NOT true about content-based filtering?

A. It can be used to predict a customer's expected rating of a new item (one without any existing ratings from other users)
B. It requires more investment in feature acquisition and engineering
C. It is a supervised learning approach
D. It is not transparent – i.e., it is not possible to say for sure why an item was recommended to a user

Word embeddings

In Word2Vec, how many vectors does the model learn for the word "model"? Note that model can have multiple meanings (a fashion model, a probabilistic model, a scaled model of a train or airplane...)

A. One vector
B. A different vector for every sentence it appears in
C. One vector for each possible meaning of the word
D. One vector for each letter in the word

Word embeddings

In Word2Vec, how many vectors does the model learn for the word "model"? Note that model can have multiple meanings (a fashion model, a probabilistic model, a scaled model of a train or airplane...)

A. One vector
B. A different vector for every sentence it appears in
C. One vector for each possible meaning of the word
D. One vector for each letter in the word

NLP

Which one is NOT the motivation behind using word embeddings instead of BOW?

A) To get the frequencies of words occurring in a document

B) To get a dense matrix representation of the words

C) To extract the contextual relationships of words

D) To understand similarities between words

NLP

Which one is NOT the motivation behind using word embeddings instead of BOW?

A) To get the frequencies of words occurring in a document

B) To get a dense matrix representation of the words

C) To extract the contextual relationships of words

D) To understand similarities between words

Multi-class Classification

To get the probability distribution of a multi-class classification, you apply:

A) Softmax function

B) Cosine similarity function

C) Sigmoid function

D) Maximum Likelihood Estimation

Multi-class Classification

To get the probability distribution of a multi-class classification, you apply:

A) Softmax function

B) Cosine similarity function

C) Sigmoid function

D) Maximum Likelihood Estimation

Neural Networks

What is the purpose of adding hidden layers in a neural network?

A. To run the network more efficiently on GPUs
B. To learn more complex patterns and features
C. To reduce training data
D. To guarantee perfect accuracy

Neural Networks

What is the purpose of adding hidden layers in a neural network?

A. To run the network more efficiently on GPUs
B. To learn more complex patterns and features
C. To reduce training data
D. To guarantee perfect accuracy

Survival Analysis

A clinical study follows patients for 12 months to see when they relapse. One patient leaves the study after 8 months without relapsing. What is this patient's data considered?

A. Complete event time
B. Missing (unusable – drop)
C. Missing (impute)
D. Right censored at 8 months

Survival Analysis

A clinical study follows patients for 12 months to see when they relapse. One patient leaves the study after 8 months without relapsing. What is this patient's data considered?

A. Complete event time
B. Missing (unusable – drop)
C. Missing (impute)
D. Right censored at 8 months