Practice Exam - Applied Machine Learning COMS W4995

Date:

Name:

UNI:

1 True/False

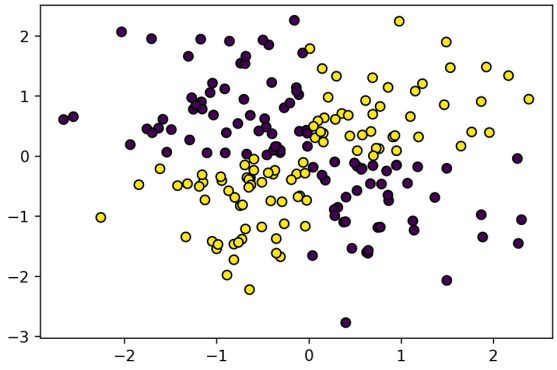
	True	False
A fast-forward git merge creates a new commit.		X
np.random.uniform() == np.random.uniform() evaluates to True.		X
Using cross-validation avoids overfitting when tuning parameters. [The question was a bit vague; cross-validation reduces the effect but doesn't prevent overfitting parameters]		X
The sign of a particular coefficients in ridge regression will be the same, no matter what the regularization parameter.		X
Stochastic gradient descent is suitable for datasets with a very high number of samples.	X	
It is good practice to standardize sparse dataset so that each feature has zero mean.		X
A node in a decision tree always contains exactly half the samples of its parent.		X
Support vector machines don't scale well to large datasets.	X	
Decision Trees are very sensitive to the scaling of the data.		X
For a perfectly calibrated classifier, 80% of the data for which p(y=1) =.8 belong to class 1.	X	

2 Multiple choice

2.1 Which of the following are non-parametric models?

- ❑ **X** Random Forest
- ❑ Linear Regression
- ❑ Logistic Regression
- ❑ **X** Nearest Neighbors
- ❑ Nearest Shrunken Centroid

2.2 Given a two-class classification dataset with the two features shown below and additional non-informative features, which of the following feature selection methods would be able to identify the these two features as informative?



- ❑ SelectPercentile(f_classif)
- ❑ **X** SelectKBest(mutual_info)
- ❑ **X** SelectFromModel(DecisionTreeClassifier())
- ❑ **X** SequentialFeatureSelector(SVC(kernel='rbf'))
- ❑ RFE(LogisticRegression())

2.3 Which of the following algorithms provide feature importances or coefficients and can be used with SelectFromModel?

- ❑ SVC(kernel="rbf")
- ❑ **X** GradientBoostingRegressor
- ❑ KNeighborsClassifier
- ❑ **X** LinearSVC

2.4 Which of the following transformations allow linear classifiers to learn non-linear decision boundaries?

- ❑ RobustScaler()
- ❑ **X** PolynomialFeatures(degree=2)
- ❑ **X** RBFSampler()
- ❑ SelectFromModel(DecisionTreeClassifier())

3 Debugging

For each code snippet, find and explain all errors given the task.

3.1 Task: Use cross-validation to assess how well feature selection and Random forest will do on the test set.

```
select = SelectPercentile(percentile=50).fit(X_train, y_train)
X_train_selected = select.transform(X_train)
scores = cross_val_score(RandomForestClassifier(n_estimators=100),
                                    X_train_selected, y_train, cv=10)
```

Feature selection should be within the cross-validation loop.

3.2 Task: Use the Box-Cox transform to preprocess data and learn a Ridge model, and visualize the coefficients. Assume that BoxCox is a scikit-learn implementation of the Box-Cox transform.

```
pipe = make_pipeline(StandardScaler(), BoxCox(), Ridge())
scores = cross_val_score(pipe, X_train, y_train, n_folds=10)
plt.barh(range(X_train.shape[0]), pipe.coef_)
```

BoxCox only works on positive data, StandardScaler will create negative entries.
Pipe has no attribute coef_, it needs to be pipe.named_steps['ridge'].coef_
there are X_train.shape[1] coefficients, not X_train.shape[0].

4 Coding

Provide code to build a LogisticRegression model and evaluate its performance on a separate test set, given a classification dataset as numpy arrays X and y.

```
X_train, X_test, y_train, y_test = train_test_split(X, y, stratify=y)
lr = LogisticRegression().fit(X_train, y_train)
score = lr.score(X_test, y_test)
```

Provide code to implement grid-searching the parameters C and gamma of an SVC in a pipeline with a StandardScaler, and evaluating the best parameter setting on a separate test set, given data as numpy arrays X and y.

```
param_grid = {'svc__C': np.logspace(-3, 3, 7), 'svc__gamma': np.logspace(-3, 3, 7) /
X.shape[0]}
X_train, X_test, y_train, y_test = train_test_split(X, y, stratify=y)
pipe = make_pipeline(StandardScaler(), SVC())
grid = GridSearchCV(pipe, param_grid, cv=5)
grid.fit(X_train, y_train)
score = grid.score(X_test, y_test)
```

5 Concepts

Answer each question with a short (2-5 sentences) explanation.

5.1 How are the “jet” and “viridis” colormaps different and why does it matter?

Viridis is conceptually uniform: the perceived difference between neighboring shades is the same everywhere in the colormap. Viridis has a constant lightness gradient, while jet varies non-linearly and non-monotonically in lightness.
The non-uniform perceived color changes in jet can create perceived edges in images where there are none, the varying lightness means that any grayscale print is completely useless.

5.2 Explain the difference between Logistic regression and linear SVMs.
Log loss vs hinge loss, logreg provides probability estimates. Bonus: all data points contribute to the solution in logreg, only some (the support vectors) in SVM.

5.3 Explain the basic idea of the RANSAC algorithm.

For linear models: draw random subset of data, fit model, compute model fit on inliers.
If enough inliers (over threshold), compare against previous best model.
Iterate, pick the model with the best fit to all inliners.

5.4 When should you use the Box-Cox transformation?
When using a linear model or other model thats sensitive to the data distribution and features have skewed distributions. Box-Cox minimizes skew, trying to create a more “Gaussian-looking” distribution.