STAT 441 Study

Juan Pablo Bello Gonzalez

September 2024

1 1. Learning Concepts

Learning:

- Classification: Categorical outcome
- Regression: Continuous outcome

Supervised: You have labels, ie right answers/confirmation for training data. Easier models than unsupervised. In unsupervised you have to make up distinctive features/categories, you don't even know how many clusters there are.
Interpretation vs. prediction tradeoff: Complex models like trees and neural netoworks are powerful at predicting but there is little interpretability to what the model coefficients actually mean.
You can restrict the class of models to allow for easy interpretation, ie trees with tunned hyperparameters. THEN optimize for prediction wihtin this class models.

1.1 Loss

MSE

1.2 Bias-Variance Tradeoff

MSE loss can be decomposed into three sources of error: **Bias, Variance, and Irreducible error**

$$E[(y_0 - \hat{f}(x_0))^2] = Var(\hat{f}(x_0))) + [Bias(\hat{f}(x_0)))]^2 + Var(\epsilon)$$

All three are non-negarive.

- Variance: How well the model generalizes to other data sets.
- Bias: how well the model fits the training set.

Overfitting: low bias high variance. Perfect fit from highly flexible model, usually with many hyperparameters, means that the model will have high variance and not predict other test sets as well as it predicts that one.
Explain the tradeoff: - Conflict of trying to simulataneously minimize the two sources of error that prevent supervised learning algorithms from generalizing beyond the training set.
MSE is the sum of three things, the graph of MSE makes a U, and the graphs of its three components are

- Irreducible error/Var(epsilon). Constant
- Variance. Increasing with flexibility. The more flexibility, the worse predictive power to data sets outside the training set the model will have.
- Bias. Decaying as flexibility increases. Higher flexibility $\implies$ lower bias $\implies$ better fit.

1.3 Bayes Classifier

A classification algorithm takes in a unit with features (a row on a table) and gives a probability distribution over the categories. Baye's classifier says, classify the unit with the most likely class.

Theorem 1 ***Conditional Bayes Error Rate****: For a given obervation x, is $1 - max_j P(Y = j|x)$ the complement of the Bayes classifier probability.*

Theorem 2 ***Overall Bayes Error Rate****: $1 - E[max_j P(Y = j|x)]$ the EXPECTATION complement of the Bayes classifier probability.*

2 2. Practical aspects

2.1 overfitting

Definition: the model fits the **random noise** of a **sample** rather than the generalizable relationship.
Occurs when the model is **too flexible**. Has too many parameters relative to the number of observations. Advanced learning algorithms are flexible, you don't need a neural network to predict y based on two features, just use some regression.

2.1.1 Defending against overfitting

- Fit to training set
- Evaluate on test set
- split must be chosen AT RANDOM. but fix seed in practice.

By splitting the data into test/train, overfitting can be avoided. If you train on 100% of the data there will be overfitting! We build a model based on the training data but evaluate the model on the test data.

2.2 cross-validation

Split data into k random subsets of equal size, called folds, then fit a model to each of the possible combinations of k-1 folds and evaluate/test on the remaining fold. Average accuracty scores to obtain a good estimate of the true predictive power of the model.
In extreme case: leave-one-out (LOO) cross validation. It's nonsense for most applications.

2.3 Evaluation measures

- Accuracy
- Sensitivity
- Specificity
- Area under the ROC curve
- F

for classification, consider:

- TP
- FP
- RN
- FN

the ordering in the name is (Correct label?, model classification label). Correct label? means that those that start with a true: TP, TN were correctly classified. The seonc instance is what that label classification was duh.
Confusion Matrix

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

2.3.1 Accuracy, Sensitivty, and Specificity

some formulas for evaluation measures

- Accuracy: (TP+TN)/N, where N is the sum of all four = #observations or sample size. Also Accuracy = $\frac{1}{n}\sum_{i=1}^{n} I(y_i = \hat{y}_i)$
- Sensitivity (TRUE POS rate): TP/(TP + FN)
- Specificity (TRUE NEG rate): TN/(TN+ FP)

Sensitivity is positive. $1 -$ Specificity is False positive rate.
Ideally we want high specificty and sensitivity.But a trade off needs to be made. The threshold of prediction... you may want to be conservative in your true predictions (high specificity low sensitiviy) if it is for Cancer diagnosis. But liberal if it is about who gets on a survival boat.

2.3.2 ROC

Area under curve is a measure of the sensitivity/specificity tradeoff.

2.3.3 F-measure

$$F = \frac{2 \times TP}{2 \times TP + FP + FN}$$

- Higher values are better.
- appropiate for 0/1 classification.

Important: - Sometimes, depending on the context, you may want to always predict false. An example: when there are ratre or uncommon events. - F measure is affected the most by this. Even if you don't predict false all the time, as long as there is a clear bias in the distribution of false classifications, your TP rate will be very low and so your F-measure will be small as well.

2.3.4 Dealing with rarely occuring categories

Not uncommon to have rare categories, ie. categories that occur infrequently.

- Macro averaging: Treat all categories the same. Default of all algorithms. Default of not doing anything in preprocessing.
- **MICRO** averaging: dominate by frequent categories, giving little weight to rare categories.

2.3.5 One-hot-encoding or "factoring" in R