DATA 100: Vitamin 10 Solutions

April 10, 2019

1 Classification

1.1 Definition

Fill in the blanks:

Classifiers are functions used to make predictions about ____ variables. They are trained on sets of ____ observations.

Select all that apply:

☑ categorical, labeled

☑ binary, labeled

□ continuous, unlabeled

□ categorical, unlabeled

Explanation: Classifiers make predictions about categorical (including binary) variables. Ther are trained on sets of labeled observations.

1.2 Consideration: Class Imbalance

Class imbalance occurs when a disproportinate amount of observations in your dataset belong to a subset of the possible classes. This can lead to classification models that achieve high accuracy, but are incapable of predicting rare outcomes. Which of the following techniques can be used to manage class imbalance?

☑ Maximize precision and recall instead of accurcy.

☑ Penalize missclassification of rare outcomes more aggresively.

□ Do nothing, typically rare outcomes aren't of interest.

☑ Use algorithms that are more robust to class imbalance.

Explanation: Class imbalance should not be ignored when training a classification model. There are many ways to handle class imbalance, including the three techniques marked above.

2 Logistic Regression

2.1 Logistic/sigmoid Function

Given a binary random variable Y and a design matrix $\mathbf{X}$, the logistic regression model assumes that $E[Y|\mathbf{X}] = P(Y = 1|\mathbf{X}) = \sigma(\mathbf{X^t}\beta)$, where $\sigma(\mathbf{X^t}\beta)$ is the logistic (or sigmoid) function. The logistic function, $\sigma(u)$, is defined as follows:

☑ $\frac{1}{1+e^{-u}}$

□ $\frac{1}{1+e^{u}}$

□ $1 + e^{-u}$

☑ $\frac{e^u}{1+e^{u}}$

□ $\frac{e^u}{1-e^{u}}$

Explanation: See the lecture slides and notebook for the definition of the logistic function.

2.2 Loss Functions

Which of the following statements are true regarding loss functions used to fit a logistic regression function?

□ The squared loss is often used to fit the parameters of a logistic regression since it is convex.

☑ The log loss is often used to fit the parameters of a logistic regression since it is convex.

□ There is a closed form solution to minimize the log loss.

☑ Log loss is also called cross-entropy loss.

Explanation: The log loss is often used to fit the parameters of logistic regression functions since it is convex. However, no closed form sollutiion exists to minimize the log loss's risk over a data set. Log loss is also known as cross-entropy loss.

2.3 Regularization

A common way to reduce the ____ of the parameter estimates of the logistic regression function is to add a (some) regularization term(s) to the loss function. However, this will increase the ____ of the estimates.

□ bias, variance

□ mean, median

☑ variance, bias

Explanation: This question relates to the bias-variance tradeoff resulting from model complexity. By regularizing the model parameters, the complexity of the logistic regression decreases. This causes the variance to decrease and the bias to increase.