Haonan Hu

2/11/2021

2863545

In-Class Problem

Problem:

- Create a Naïve Bayesian Classifier for the iris dataset.
- Given:
  - The iris data set contains 150 samples of data, 50 for each variety of iris: Iris-setosa, Iris-versicolor, & Iris-virginica
  - We will use 149 samples of the data to train the classifier, and test it with one sample of Iris-virginica which has the following features:
    - sepal-length = 5.9
    - sepal-width = 3
    - petal-length = 5.1
    - petal-width = 1.8

1. Give the formula for the posterior numerator for each variety, e.g., posterior numerator(Iris-setosa).
2. Calculate P for each variety, e.g., P(Iris-setosa)
3. Give the formula for p(sepal-length|Iris-setosa), if the mean value and variance of sepal-length for Iris-setosa is 5.0 and 0.12, respectively. Substitute the values for x, μ, and σ^2 into the formula.
4. How many conditional probabilities will the Naïve Bayesian Classifier need to calculate to classify the test sample?
5. If posterior numerator(Iris-setosa) = 0.005, posterior numerator(Iris-versicolor) = 0.002, and posterior numerator(Iris-virginica) = 0.003, which variety did the Naïve Bayesian Classifier predict the test sample to be?

1. Give the formula for the posterior numerator for each variety

Posterior numerator(iris-setosa) = P (iris-setosa) * P (sepal-length| iris-setosa) * P (sepal-width| iris-setosa) * P (petal-length| iris-setosa) * P (petal-width| iris-setosa)

Posterior numerator(iris-versicolor) = P (iris-versicolor) * P (sepal-length| iris-versicolor) * P (sepal-width| iris-versicolor) * P (petal-length| iris-s versicolor) * P (petal-width| iris-versicolor)

Posterior numerator(iris-virginica) = P (iris- virginica) * P (sepal-length| iris-virginica) * P (sepal-width| iris- virginica) * P (petal-length| iris- virginica) * P (petal-width| iris-virginica)

2. **Calculate P for each variety**

   P(iris-setosa) = 50 / 149 ~= 0.3356
   P(iris-versicolor) = 50 /149 ~= 0.3356
   P(iris-virginica) = 49 / 149 ~= 0.3289

3. **Give the formula for P (sepal-length | iris-setosa) if the mean value and variance of sepal length for iris-setosa is 5 and 0.12. Substitute the values for x, u, and variance into the formula**

   P (sepal-length | iris-setosa) = $\frac{1}{\sqrt{2*\pi*0.12}} * e^{-\frac{(5.9-5)^2}{2*0.12}}$

4. **How many conditional probabilities will the Naïve Bayesian Classifier need to calculate to classify the test samples?**

   3(varieties) * 4(features) = 12 conditional probabilities

5. **If posterior numerator(iris-setosa) = 0.005, posterior numerator(iris-versicolor) = 0.002 and posterior numerator(iris-virginica) = 0.003, which variety did the Naïve Bayesian Classifier predict the test sample to be.**

   0.005 > 0.003 > 0.002, so we conclude that the variety of test sample is iris-setosa(0.005).