2021 Introduction to Machine Learning Program Assignment #1 - Naïve Bayes

TA: Evan Chang toosyou.tw@gmail.com (mailto:toosyou.tw@gmail.com)

This programming assignment aims to help you understand the algorithm behind Naïve Bayes classifier and the basic workflow of machine learning.

Before we start

> Join the discord server for TA support (https://discord.gg/XJkvmNrcjp)
>
> - Ask questions on it, and we shall reply.
> - Try not to ask for obvious answers or bug fixes.
> - Memes and chit-chat welcome.

Objective

There are two datasets that need to be analyzed. For each dataset, you have to do the following:

1. Data Input - 5%
2. Data Visualization - 15%
   - For mushroom dataset
     - Show the data distribution by **value frequency** of every feature.
   - For Iris dataset
     - Show the data distribution by **average, standard deviation, and value frequency(binning might be needed)** of every feature.
   - Split data based on their labels (targets) and show the data distribution of each feature again.
3. Data Preprocessing - 5% + (10%)
   - Drop **features** with any missing value.
   - Transform data format and shape so your model can process them.
   - **Shuffle the data**.
   - Bonus: any other transformation boosts the final performance. - (10%)
4. Model Construction - 20%
   - You must construct two Naïve Bayes classifiers for the two datasets.
     - **You may use any package avaliable** as long as the classifiers fit the following description.

- Naïve Bayes divider M in log-space:
    - $M(\mathbf{q}) = \underset{Y\in\mathbb{T}}{\operatorname{argmax}}[\log P(Y) + \Sigma_{i=1}^{m} \log P(X_i|Y)]$
        - where $\mathbf{q} = \{X_1, X_2, ..., X_m\}$ is a sample to be predicted, whose features are X_1 to X_m. $\mathbb{T}$ is the set of all possible labels.
- For the mushroom dataset, whose features are all **categorical**, $P(X_i|Y)$ must be computed with and without Laplace smoothing for result comparison. - 10%
    - Without Laplace smoothing
        - $P(X_i|Y) = \frac{N(X_i|Y)}{N(Y)}$
    - Laplace smoothing
        - $P(X_i|Y) = \frac{N(X_i|Y)+k}{N(Y)+k\tau}$
            - where τ is the number of all possible events of feature X_i
- For Iris dataset, whose features are all **numerical**, assume $P(X_i|Y)$ follows a 1D-Normal(Gaussian) distribution. - 10%
    - $P(X_i|Y) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
        - where μ, σ are the mean and standard deviation of feature X_i respectively, while label Y is determined.

5. Train-Test-Split - 10%
    - Two validation methods need to be implemented.
        1. Holdout validation (https://en.wikipedia.org/wiki/Cross-validation_(statistics)#Holdout_method) with the ratio $7:3$
        2. K-fold cross-validation (https://en.wikipedia.org/wiki/Cross-validation_(statistics)#k-fold_cross-validation) with $K = 3$
            - Obtain the final performance by **averaging** all folds' performance.
6. Results - 20%
    - Obtain the performances of all experiment settings **in tables** by the following metrics:
        1. Confusion matrix
        2. Accuracy
        3. Sensitivity(Recall)
        4. Precision
7. Comparison & Conclusion - 5%
8. Questions - 25%
    - For the mushroom dataset
        1. Show $P(X_{stalk-color-below-ring}|Y = e)$ **with** and **without** Laplace smoothing by bar charts - 10%
    - For Iris dataset
        1. What are the values of μ and σ of assumed $P(X_{petal\_length}|Y = \text{Iris Versicolour})$? - 5%

2. Use a graph to show the probability density function of assumed $P(X_{petal_length}|Y = \text{Iris Versicolour})$10%

Data

1. Mushroom dataset

- Data can be downloaded here:
    - https://archive.ics.uci.edu/ml/datasets/mushroom (https://archive.ics.uci.edu/ml/datasets/mushroom)
- Please **NOTE** that the first column is the label (edible=e, poisonous=p)
- Data Set Information
    - This data set includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family (pp. 500-525). Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter class was combined with the poisonous one. The Guide clearly states that there is no simple rule for determining the edibility of a mushroom; no rule like ‘‘leaflets three, let it be’’ for Poisonous Oak and Ivy.
- Attribute Information
    1. cap-shape: bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s
    2. cap-surface: fibrous=f,grooves=g,scaly=y,smooth=s
    3. cap-color: brown=n,buff=b,cinnamon=c,gray=g,green=r, pink=p,purple=u,red=e,white=w,yellow=y
    4. bruises?: bruises=t,no=f
    5. odor: almond=a,anise=l,creosote=c,fishy=y,foul=f, musty=m,none=n,pungent=p,spicy=s
    6. gill-attachment: attached=a,descending=d,free=f,notched=n
    7. gill-spacing: close=c,crowded=w,distant=d
    8. gill-size: broad=b,narrow=n
    9. gill-color: black=k,brown=n,buff=b,chocolate=h,gray=g, green=r,orange=o,pink=p,purple=u,red=e, white=w,yellow=y
    10. stalk-shape: enlarging=e,tapering=t
    11. stalk-root: bulbous=b,club=c,cup=u,equal=e, rhizomorphs=z,rooted=r, **missing=?**
    12. stalk-surface-above-ring: fibrous=f,scaly=y,silky=k,smooth=s
    13. stalk-surface-below-ring: fibrous=f,scaly=y,silky=k,smooth=s
    14. stalk-color-above-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y
    15. stalk-color-below-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y

16. veil-type: partial=p,universal=u
17. veil-color: brown=n,orange=o,white=w,yellow=y
18. ring-number: none=n,one=o,two=t
19. ring-type: cobwebby=c,evanescent=e,flaring=f,large=l, none=n,pendant=p,sheathing=s,zone=z
20. spore-print-color: black=k,brown=n,buff=b,chocolate=h,green=r, orange=o,purple=u,white=w,yellow=y
21. population: abundant=a,clustered=c,numerous=n, scattered=s,several=v,solitary=y
22. habitat: grasses=g,leaves=l,meadows=m,paths=p, urban=u,waste=w,woods=d

2. Iris dataset

- Data can be downloaded here:
  - https://archive.ics.uci.edu/ml/datasets/iris (https://archive.ics.uci.edu/ml/datasets/iris)
- Data Set Information
  - This is perhaps the best known database to be found in the pattern recognition literature. Fisher's paper is a classic in the field and is referenced frequently to this day. (See Duda & Hart, for example.) The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other. Predicted attribute: class of iris plant. This is an exceedingly simple domain.
- Attribute Information
  1. sepal length in cm
  2. sepal width in cm
  3. petal length in cm
  4. petal width in cm
  5. class:
     - Iris Setosa
     - Iris Versicolour
     - Iris Virginica

Submission & Scoring Policy

- Please submit a **zip** file, which contains the following, to the newE3 system.
  1. Report
     - Explanation of how your code works.
     - All the content mentioned above.
     - Your name and student ID at the very beginning - 10%
     - Accept formats: **HTML**
       - From markdowns (https://hackmd.io/) or jupyter notebooks.

    2. Source codes
        - Accept languages: **python3**
        - Accept formats: **.ipynb** (https://jupyter.org/)
        - **Package-provided models are allowed**
- Your score will be determined mainly by the submitted report.
    - If there’s any problem with your code, TA might ask you (through email) to demo it. Otherwise, no demo is needed.
- Scores will be adjusted at the end of the semester for them to fit the school regulations.
- **Plagiarizing is not allowed**.
    - Plagiarizing is checked by MOSS (https://theory.stanford.edu/~aiken/moss/) and manually afterward.
    - You will get **ZERO** on that homework if you get caught the first time.
    - The second time, you’ll **FAIL** this class.
    - **抄襲第一次作業零分，第二次當掉**

Acknowledgments

- Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml (http://archive.ics.uci.edu/ml)]. Irvine, CA: University of California, School of Information and Computer Science.

Tools that might be useful

- Jupyter Lab (https://jupyter.org/) - Better data science experience
- numpy (https://numpy.org/) - Math thingy
- matplotlib (https://matplotlib.org/tutorials/introductory/pyplot.html) - Plot thingy
- pandas (https://pandas.pydata.org/) - Data thingy
- scikit-learn (https://scikit-learn.org/stable/) - Machine Learning and stuff