Project 2: Price assurance?

DS 3010 Data Science III: Computational Data Intelligence

Aukkawut Ammartayakun

April 19, 2022

Worcester Polytechnic Institute

Outline

1. Introduction

2. Case study: Mobile Phone

3. Accuracy Problem

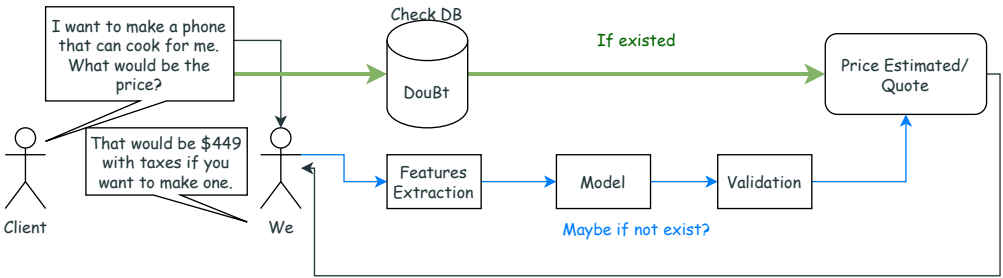
   Structured vs Quasi-Continuous

   n-Binary vs Dirichlet?

4. Conclusion

Introduction

Introduction



Case Study: Mobile Phone

Dataset

Numerical information

- Battery power
- Talk time

Dataset

Numerical information

- Battery power
- Talk time
- Clock speed?
- Front camera and primary camera pixel count?
- Internal storage and RAM?
- Mobile dimensions?

Dataset

Numerical information

- Battery power
- Talk time
- Clock speed?
- Front camera and primary camera pixel count?
- Internal storage and RAM?
- Mobile dimensions?

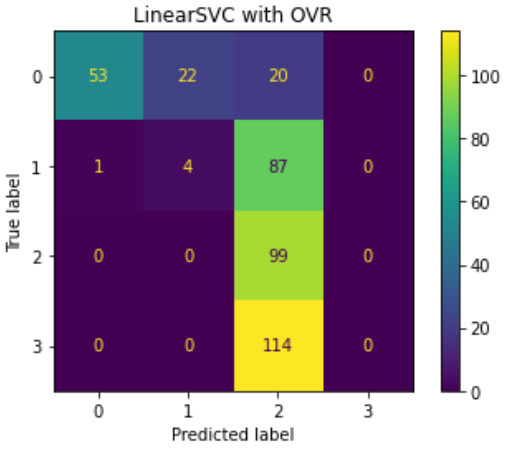
Non-numerical information

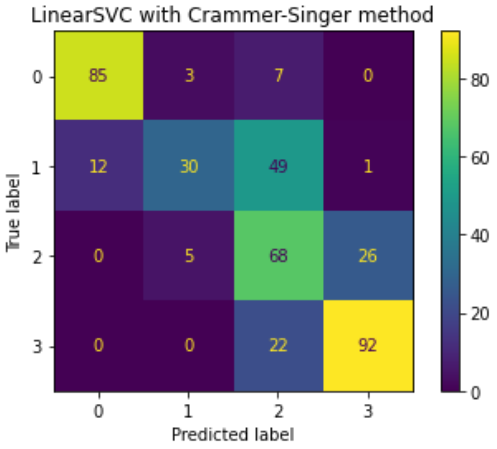
- Connectivity (Bluetooth,4G,etc.)
- Touchscreen

Modeling

- Linear model
  - Linear Support Vector Classification (LinearSVC)
  - Logistic Regression (LR)
- Non-linear model
  - Multi-Layer Perceptrons (MLP)

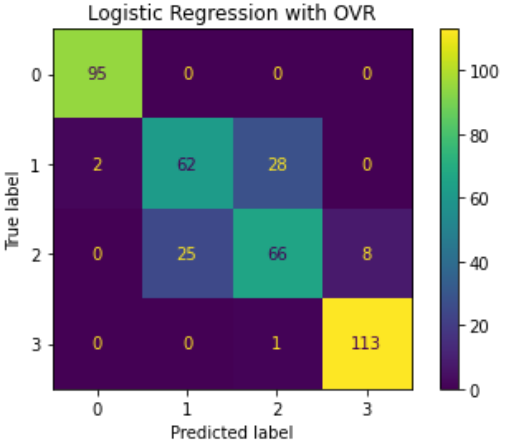
LinearSVC

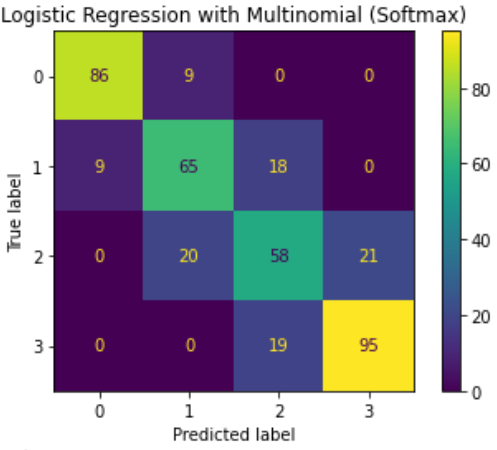




Accuracy=0.39Accuracy=0.68

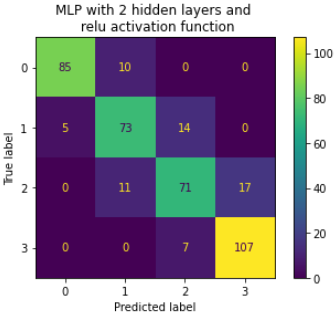
Logistic Regression

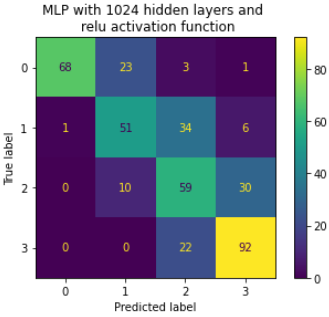


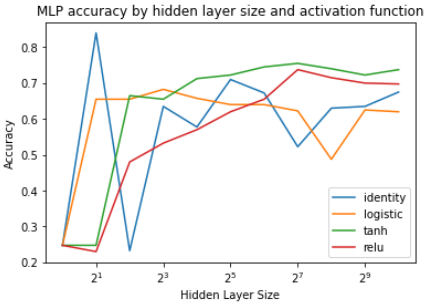


Accuracy = 0.84Accuracy =0.76

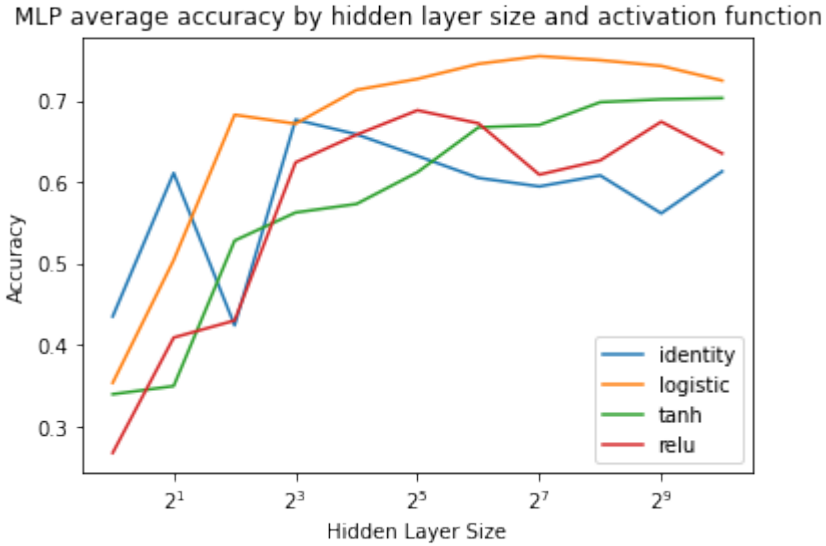
Multi-Layer Perceptrons







Multi-Layer Perceptrons



Accuracy Problem

Why ordinal classification instead of normal classification?

Let say we have observations

$$X_1, \ldots, X_n \sim \text{Categorical}(|\mathcal{C}|, \rho)$$

with a class set $\mathcal{C} = \{c_1, c_2, \ldots, c_k\}$ and probability measure ρ.

Why ordinal classification instead of normal classification?

Let say we have observations

$$X_1, \ldots, X_n \sim \text{Categorical}(|\mathcal{C}|, \rho)$$

with a class set $\mathcal{C} = \{c_1, c_2, \ldots, c_k\}$ and probability measure ρ.

Our classification task is to find $\hat{q} : \mathcal{X} \to \mathcal{C}$ such that their classes are **closed** to the real classes.

Why ordinal classification instead of normal classification?

Let say we have observations

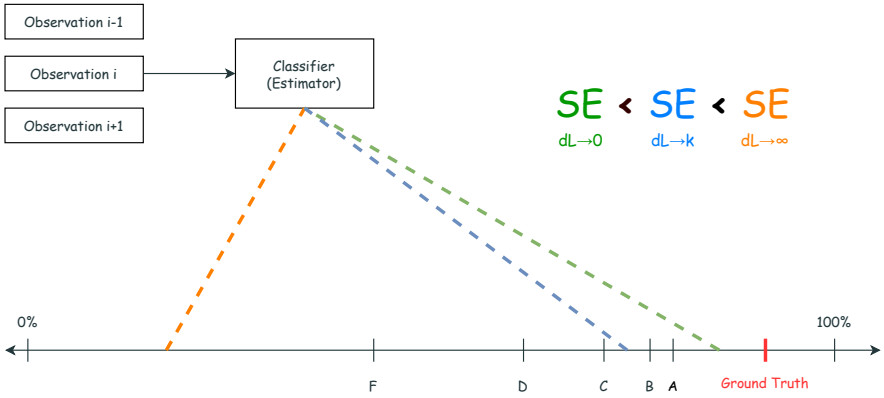
$$X_1, \ldots, X_n \sim \text{Categorical}(|\mathcal{C}|, \rho)$$

with a class set $\mathcal{C} = \{c_1, c_2, \ldots, c_k\}$ and probability measure ρ.

Our classification task is to find $\hat{q} : \mathcal{X} \to \mathcal{C}$ such that their classes are **closed** to the real classes.

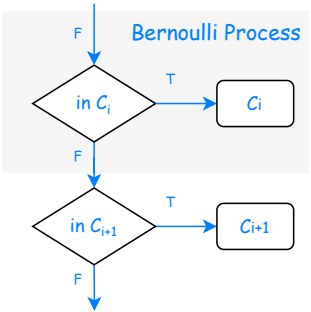
Generally, we would use categorical cross-entropy as our objective. However, what if the class set is quasi-continuous in such the way that distance measure can be used?

Why ordinal classification instead of normal classification?

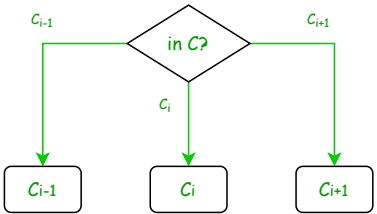


In this scenario, the quasi-continuous space can be defined as discretized space. Notice that the continuity implies that the blue prediction should have less penalty to the objective function compared to orange one.

One-vs-Rest vs Multinomial?



One-VS-Rest



Multinomial

Conclusion

Conclusion

Conclusion

- Will our business possible?

Conclusion

- Will our business possible?
  - We have close to all the building blocks!

Conclusion

- Will our business possible?
  - We have close to all the building blocks!
- How can we deal with unknown data like ”phone that can cook you a food”?

Conclusion

- Will our business possible?
  - We have close to all the building blocks!
- How can we deal with unknown data like "phone that can cook you a food"?
  - Zero-shot learning with NLP knowledge!

Conclusion

- Will our business possible?
  - We have close to all the building blocks!
- How can we deal with unknown data like "phone that can cook you a food"?
  - Zero-shot learning with NLP knowledge!
- How can we deal with newly released product?
  - Online learning!