# Naïve Bayes Leaner for `Adult` Database

**COMP30027 2022 Assignment 1**

Ni Ding

26 Mar 2022

# Purpose of Project

## Developing Engineering Skills: Machine Learning (ML) Capabilities

- **Problem Solving** (build Naïve Bayes Learner)
  - *(Data) Analytics*: read and understand assignment spec and `adult.csv`
  - *Computer Modeling*: build and evaluate a Naïve Bayes Learner, How?

    Python coding? Yes, but more ML skills

  - *Troubleshooting* (Analytical Skills): guided by questions Q1 to Q4 (Q1&2 for individual), beyond code debugging
    - ⁎ Interpret and explain the results: expected/unexpected, good/bad, .....
    - ⁎ More Importantly, how.....why..... (other choices) $\Longrightarrow$ Improvement
  - *Communication skills*: disseminate your discovery/innovation (to us), clean code (w/ proper comments), concise answers, ...

# (Data) Analytics

**Database:** `Adult.csv`: 11 attributes, a binary class label, 1,000 records

| age | work class | education | education num | ... | label |
|---|---|---|---|---|---|
| 68 | ? | 1st-4th | 2 | ... | $<=$ 50K |
| 39 | State-gov | Bachelors | 13 | ... | $<=$ 50K |
| 50 | Self-emp-not-inc | Bachelors | 13 | ... | $<=$ 50K |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |

**Assignment Spec/Project Description** (excl. question part): Naïve Bayes Learner,

$$\hat{c} = \arg\max_{c_j} P(c_j) \prod_{i=1}^{m} P(X_i|c_j)$$

to be implemented and evaluated by `preprocess()`, `train()`, `test()` and `evaluate()`

# Computer Modeling

**Training** `train()`: using attribute values $(a_1, \ldots, a_m)$'s and their labels $c_j$'s in training set to lean Naïve Bayes Learner $f$ (a formal expression):

$$\begin{aligned}\hat{c} &= f(a_1, \ldots, a_m) \\ &= \arg\max_{c_j} \Pr(C = c_j) \prod_{i=1}^{m} \Pr(X_i = a_i | C = c_j)\end{aligned}$$

$m = 11$, attribute $X_i$ (e.g., 'age', 'education'), label $C$ ($c_1 =$ '$<= 50$K',$c_2 =$ '$> 50$K')

- kowning all parameters that determine $f$: prior $\Pr(C = c_j)$ and conditional probs. $\Pr(X_i = a_i | C = c_j)$ for all attribute values $a_i$'s and $c_j$'s, for each attribute $X_i$
- a mixture of numeric and nomial attributes
- $\Pr(X_i = a_i | C = c_j)$: numeric (Gaussian in four functions and KDE in Q2(a)) and nominal attributes, refer to (Thu lecture in week 2 and Tue lecture in week 3)
- issues/concerns: data structure, missing values (see spec, e.g., Q3), .....

# Computer Modeling

**Testing** `test()`: using attribute values $(t_1, \ldots, t_m)$'s in testing set to get $\hat{c}$:

$$\begin{aligned}\hat{c} &= f(t_1, \ldots, t_m) \\ &= \arg\max_{c_j} \Pr(C = c_j) \prod_{i=1}^{m} \Pr(X_i = t_i | C = c_j)\end{aligned}$$

issues/concerns? e.g., missing values: decide by yourself that makes the most sense.

**Evaluating** `evaluate()`: comparring true label $c^*$ and predicted label $\hat{c}$ to get accuracy, confusion matrix

| | | **Predicted** | |
|---|---|---|---|
| | | Positive | Negative |
| **True** | Positive | TP | FN |
| | Negative | FP | TN |

and F1 score (Thu lecture in week 3)

# Troubleshooting

**Q1: interpreting, analyzing and reasoning the output of `evaluate()`**

## Keyword Extraction

Sensitivity and specificity......should have both sensitivity and specificity high......calculate the sensitivity and specificity......see a difference......what causes this difference......suggestions to improve the model performance......

- Expectation ⟨ red ⟩
- To do ⟨ purple ⟩
  - Observation: values of sensitivity and specificity, (mis)align w/ expectation?
  - Reasoning: why the observation, thinking of possibilities and justify
  - Suggestion: based on your reasoning. $\Longrightarrow$ completes troubleshooting!

# Troubleshooting

**Q2: exploring other choices for `train()` and `test()`**

## Sub-task (a)

Gaussian vs. KDE for modeling $P(X_i|c_j)$, aka $\Pr(X_i = a_i|C = c_j)$, for numeric $X_i$

- Suitability of two probabilistic models for each numeric $X_i$: **Justify**

## Sub-task (b)

Different $m$ in $m$-fold cross-validation: $m \in \{2, 10\}$

- Observation and Interpreting (some reasoning): the changes in evaluation metrics (e.g., accuracy, see spec) for different $m$ and over all folds, why difference (Thu lecture in week 3)
- Conclusion: summerize and answer the quesion

# Troubleshooting

**Q3: exploring different methods for handling missing values for nominal $X_i$**

## Options

1. Missing value '?' $\Longrightarrow$ new category
2. Ignore missing value: <u>how to ignore</u>

▲ Try both large and small datasets; <u>what you should look at to compare</u>

▲ Observation-conclusion (incl. reasoning) approach: justify!

# Troubleshooting

**Q4: exploring how to use information gain (IG) and gain ratio (GR) in Naïve Bayes (in fact beyond) The most Interesting part!**

## Sub-task (a)

Obtain GR: attributes vs. class

- throwing out attributes $X_i$'s one by one according to ascending order of GRs

▲ Observe the change of accuracy; justify and reason the observatons.

## Sub-task (b)

Obtain IG: for each pair of attributes $X_i$ and $X_{i'}$, for all $i, i' \in \{1, \ldots, m\}$ such that $i \neq i'$

- predict one by another.

▲ Justify, alwasy justify!......

# Finally

Good Luck!