# Homework 1
# Probability, MLE, MAP, KNN and Naive Bayes[1]

CMU 10-701: Introduction to Machine Learning (Fall 2021)

piazza.com/cmu/fall2021/10701/home
OUT: Wednesday, Sep 8th, 2021

DUE: Wednesday, Sep 22th, 2021, 11:59pm

## START HERE: Instructions

- **Collaboration policy:** Collaboration on solving the homework is allowed, after you have thought about the problems on your own. It is also OK to get clarification (but not solutions) from books or online resources, again after you have thought about the problems on your own. There are two requirements: first, cite your collaborators fully and completely (e.g., "Jane explained to me what is asked in Question 2.1"). Second, write your solution *independently*: close the book and all of your notes, and send collaborators out of the room, so that the solution comes from you only. See the Academic Integrity Section in our course syllabus for more information: https://piazza.com/class_profile/get_resource/ksetdrgdkob78/ksqc9bxxjt56ic
- **Late Submission Policy:** See the late submission policy here: https://piazza.com/class_profile/get_resource/ksetdrgdkob78/ksqc9bxxjt56ic
- **Submitting your work:**
  - **Gradescope:** There will be two submission slots for this homework on Gradescope: Written and Programming.
    For the written problems such as short answer, multiple choice, derivations, proofs, or plots, we will be using the written submission slot. Please use the provided template. The best way to format your homework is by using the Latex template released in the handout and writing your solutions in Latex. However submissions can be handwritten onto the template, but should be labeled and clearly legible. If your writing is not legible, you will not be awarded marks. Each derivation/proof should be completed in the boxes provided below the question, **you should not move or change the sizes of these boxes** as Gradescope is expecting your solved homework PDF to match the template on Gradescope. If you find you need more space than the box provides you should consider cutting your solution down to its relevant parts, if you see no way to do this, please add an additional page at the end of the homework and guide us there with a 'See page xx for the rest of the solution'.
    You are also required to upload your code, which you wrote to solve the final question of this homework, to the Programming submission slot. Your code may be run by TAs so please make sure it is in a workable state.
    Regrade requests can be made after the homework grades are released, however this gives the TA the opportunity to regrade your entire paper, meaning if additional mistakes are found then points will be deducted.

For multiple choice or select all that apply questions, shade in the box or circle in the template document corresponding to the correct answer(s) for each of the questions. For LaTeXusers, use ■ and ●for shaded boxes and circles, and don't change anything else. If an answer box is included for showing work, **you must show your work**!

[1]Compiled on Wednesday 8th September, 2021 at 18:14

# 1 Probability Review [15pts]

A group of travellers find themselves lost in a cave. They come upon 3 tunnels $A$, $B$, $C$. Both tunnels $A$ and $B$ are closed loops that do not lead to an exit and in fact lead right back to the entrance of the 3 tunnels. Tunnel $C$ is the tunnel which leads to the exit. If they go through tunnel $A$, then it takes 2 days to go through the tunnel. If they go through tunnel $B$, then it take 1 day to go through the tunnel. If they go through tunnel $C$, then they immediately leave the cave. Suppose the travellers choose tunnels $A$, $B$ and $C$ with constant probability 0.6, 0.3, 0.1 every time. (For the following questions please round your answer up to 4 digits.)

1. **[6 pts]** Suppose we record down the travellers choices into a sequence (e.g., $ABBA\ldots C$). What is the probability that the pattern $AAB$ appears in the sequence before any $BAA$ appears?
   **Note:** You should also count cases where $AAB$ appears in the sequence and $BAA$ does not.

**Final Answer**

0.27

**Work for Final Answer**

The pattern AAB appears before BAA = no B before AAB appears

⇒ The overall pattern should be AAA.....AAB

$P(AAB) + P(AAAB) + P(AAAAB) + \cdots\cdots$

$= P(AAB) + P(A)P(AAB) + P(A)P(A)P(AAB) + \cdots$

$= P(AAB)\left(1 + P(A) + [P(A)]^2 + \cdots + [P(A)]^{\infty}\right)$

$= P(A)P(A)P(B)\left[1 + P(A) + P(A)^2 + \cdots + P(A)^{\infty}\right]$

$= 0.6 \times 0.6 \times 0.3 \times (1 + 0.6 + 0.6^2 + \cdots + 0.6^{\infty}) = 0.108 \times \frac{1}{1-0.6} = \frac{0.108}{0.4} = 0.27$ ✱

2. [3 pts] What is the expected number of days that the travellers will be lost in the cave?

**Final Answer**

15

**Work for Final Answer**

The expected days wasted for one wrong choice is

$P(\text{Not C}) \times \left[\left(\frac{P(A)}{1-P(C)} \times 2\text{days}\right) + \left(\frac{P(B)}{1-P(C)} \times 1\text{ day}\right)\right] = 0.9 \times \left[\frac{0.6}{0.9} \times 2 + \frac{0.3}{0.9} \times 1\right] = 0.9 \times \frac{5}{3}$

n' Wrong choices and 1 right choice in the end

$= \left(\sum_{n=1}^{\infty} n \times \frac{5}{3} \times (0.9)^n\right) \times 0.1 = 0.1 \times \frac{5}{3} \times \sum_{n=1}^{\infty} n(0.9)^n$

$= 0.1 \times \frac{5}{3} \times \frac{0.9}{(1-0.9)^2} = \frac{5}{3} \times 9 = \frac{45}{3} = 15$ #

3. [6 pts] What is the variance of days that the travellers will be lost in the cave? (Hint: To compute $Var(T)$ for a random variable $T$, you can either compute $E[T^2]$ first and then $Var(T)$ or directly compute the variance using the law of total variance.)

**Final Answer**

250

**Work for Final Answer**

$Var(T) = E[T^2] - E[T]^2$

$E[T]^2 = 15^2 = 225$

$E[T^2] = \sum P(T)T^2$

$= 0.1 \times \left(\frac{5}{3}\right)^2 \times \sum_{n=1}^{\infty} n^2(0.9)^n = 0.1 \times \frac{25}{9} \times 0.9 \times \frac{1+0.9}{(1-0.9)^3} = 475$

$\Rightarrow Var(T) = 475 - 225 = 250$ #

## 2 MLE and MAP [25pts]

Please note that for this section (MLE and MAP) you do not need to show proof that the optimum point is a maximum.

### 2.1 MLE with Exponential Family

likelihood

1. [7pts] The exponential family of distributions has the form $P(x \mid \theta^*) = h(x) \exp[\theta^* \cdot \phi(x) - A(\theta^*)]$. It may look unfamiliar, but in fact many well-known distributions including Gaussian, Bernoulli, Geometric and Laplace distributions belong to this family[2]. Suppose we are given $n$ i.i.d samples $X = \{x^1, x^2, \ldots, x^n\}$ drawn from the distribution $P(x \mid \theta^*)$. Derive the Maximum Likelihood Estimator $\widehat{\theta}_{\text{MLE}}$ for this true parameter $\theta^*$.

   Here you can assume that $A$ is convex and differentiable, and that the derivative $A'$ is invertible. Your answer should be in terms of $A$, $\phi(x)$ and $h(x)$.

**Final Answer**

$$\frac{1}{n}(A')^{-1}\sum_{x=1}^{n}\phi(x)$$

**Work for Final Answer**

$$P(x|\theta^*) = h(x)\exp[\theta^* \cdot \phi(x) - A(\theta^*)]$$

$$\text{MLE}: \hat{\theta}_{MLE} = \arg\max_{\theta^*} P(x|\theta^*)$$

$$= \arg\max_{\theta^*} \prod_{i}^{n} P(x|\theta^*)$$

$$\log L(\hat{\theta}) = \arg\max_{\theta^*} \log \prod_{x=1}^{n} P(x|\theta^*)$$

$$= \sum_{x=1}^{n} \log\left(h(x) \cdot e^{\theta^* \cdot \phi(x) - A(\theta^*)}\right)$$

$$= \sum_{x=1}^{n} [\log h(x) + \log e^{\theta^* \cdot \phi(x) - A(\theta^*)}]$$

$$= \sum_{x=1}^{n} [\log h(x) + \theta^* \cdot \phi(x) - A(\theta^*)]$$

$$\Rightarrow \frac{dL(\theta)}{d\theta} = \sum_{x=1}^{n} (\phi(x) - A'(\theta^*)) = \sum_{x=1}^{n} \phi(x) - nA'(\theta^*) = 0$$

$$\Rightarrow \sum_{x=1}^{n} \phi(x) = nA'(\theta^*)$$

$$\Rightarrow \hat{\theta}_{MLE} = \frac{(A')^{-1}\sum_{x=1}^{n}\phi(x)}{n}$$

[2] To see the parameter setting for each of these distributions, which makes them become special cases of exponential distributions you can check https://en.wikipedia.org/wiki/Exponential_family#Table_of_distributions.

## 2.2 MLE and MAP with Pareto Distribution

1. **[5 pts]** The Pareto distribution has the form

$$f(x) = \frac{ab^a}{x^{a+1}}, \quad x \geq b$$

with the parameters $a, b > 0$. For our purposes, assume that $b$ is known. We obtain $n$ i.i.d. data points $x^1, x^2, \ldots, x^n$ from the Pareto distribution. Find the MLE estimate $\hat{a}$.

**Final Answer**

$$\frac{n}{\sum_{i=1}^{n} \log x_i - n \log b}$$

**Work for Final Answer**

$$L(a) = \prod_{i=1}^{n} \frac{ab^a}{x_i^{a+1}} = (ab^a)^n \times \prod_{i=1}^{n} x_i^{-(a+1)}$$

$$\Rightarrow \log L(a) = n \log ab^a + \sum_{i=1}^{n} \log x_i^{-(a+1)}$$

$$= n(\log a + a \log b) - (a+1) \sum_{i=1}^{n} \log x_i$$

$$\Rightarrow \frac{d \log L(a)}{da} = n \cdot \frac{1}{a} + n \log b - \sum_{i=1}^{n} \log x_i = 0$$

$$\frac{n}{a} = \sum_{i=1}^{n} \log x_i - n \log b$$

$$a = \frac{n}{\sum_{i=1}^{n} \log x_i - n \log b}$$

2. **[10 pts]** Now suppose $a$ has a Gamma$(\alpha, \beta)$ prior distribution with probability density function:

$$f(a) = \frac{\beta^\alpha}{\Gamma(\alpha)} a^{\alpha-1} e^{-\beta a}$$

The parameters $\alpha > 0, \beta > 0$ are both known. Find the posterior distribution of $a$ given $x$, and find the MAP estimate $\tilde{a}$.

**Posterior Distribution**

$$\prod_{i=1}^{n} \frac{a b^a}{x_i^{a+1}} \times \frac{\beta^\alpha}{\Gamma(\alpha)} a^{\alpha-1} e^{-\beta a}$$

**MAP estimate $\tilde{a}$**

$$\frac{\alpha - 1 + n}{\beta - n \log b + \sum_{i=1}^{n} \log x_i}$$

**Work for Final Answer**

$$P(a|x) = P(x|a) P(a)$$

$$= \prod_{i=1}^{n} \frac{a b^a}{x_i^{a+1}} \times \frac{\beta^\alpha}{\Gamma(\alpha)} a^{\alpha-1} e^{-\beta a}$$

$$= \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot e^{-\beta a} \cdot a^{\alpha-1} \cdot \prod_{i=1}^{n} \frac{a b^a}{x^{a+1}}$$

$$= \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot e^{-\beta a} \cdot a^{\alpha-1} \cdot a^n \cdot \prod_{i=1}^{n} \frac{b^a}{x^{a+1}}$$

$$\xrightarrow{\log} \alpha \log \beta - \log \Gamma(\alpha) + \log e^{-\beta a} + \log a^{\alpha-1+n} + \sum a \log b - (a+1) \sum \log x$$

$$= \alpha \log \beta - \log \Gamma(\alpha) - \beta a + (\alpha - 1 + n) \log a + a \sum \log b - (a+1) \sum \log x$$

$$\frac{d \log P(a|x)}{da} = -\beta + \frac{\alpha - 1 + n}{a} + \sum (\log b - \log x) = 0$$

$$\Rightarrow \frac{\alpha - 1 + n}{a} = \beta - \sum (\log b - \log x)$$

$$a = \frac{\alpha - 1 + n}{\beta - \sum (\log b - \log x)} = \frac{\alpha - 1 + n}{\beta - n \log b + \sum \log x}$$

3. **[3 pts]** Assume $\sum_{i=1}^{n} \log \frac{x^i}{b} \to \infty$ as $n \to \infty$ for the Pareto distribution. Compare the MLE ( $\hat{a}$ ) and the MAP ( $\tilde{a}$ ) as $n \to \infty$ and describe your findings.

$$\sum_{i=1}^{n} \log \frac{x_i}{b} = \sum_{i=1}^{n} (\log x_i - \log b) = \sum_{i=1}^{n} \log x_i - n \log b$$

$$\text{MLE}(\hat{a}) = \frac{n}{\sum_{i=1}^{n} \log x_i - n \log b} \xrightarrow{\frac{d}{dn}} \frac{1}{-\log b}$$

$$\text{MAP}(\tilde{a}) = \frac{\alpha + n - 1}{\beta - n \log b + \sum_{i=1}^{n} \log x_i} \xrightarrow{\frac{d}{dn}} \frac{1}{-\log b}$$

Same

The more data points we have for MAP, the less the effect of prior to the posterior.
=> "washed out"

## 3 K-Nearest Neighbors: Black Box [10 Points]

1. **[6 pts]** In a KNN classification problem, assume that the distance measure is not explicitly specified to you. Instead, you are given a "black box" where you input a set of instances $P_1, P_2, \ldots P_n$ and a new example $Q$, and the black box outputs the nearest neighbor of $Q$, say $P_i$ and its corresponding class label $C_i$. Is it possible to construct a KNN classification algorithm (w.r.t the unknown distance metrics) based on this black box alone? If so, how and if not, why not?

Wcq* uc a_l glnsr rfc q_kc O dmp i rgkcq* _lb dmp c_af rgkc uc pckmtc rfc msrnsr $N_g$ dpmk rfc glqr_lacq _lb pcampb rfc j_`cj $A_g$, Rfcl uc ugjj f_tc $A_/$* $A_0$* „* $A_i$ gl rfc mpbcp md lc_pcqr rm rfc $i^{rf}$ lc_pcqr, Dgl_jjw* uc a_l aj_qqgdw O _q rfc aj_qq rf_r f_q rfc fgefcqr _kmslrgl $A_/$* $A_0$* „* $A_i$,

2. **[4 pts]** If the black box returns the $j$ nearest neighbors (and their corresponding class labels) instead of the single most nearest neighbor (assume $j \neq k$), is it possible to construct a KNN classification algorithm based on the black box? If so how, and if not why not?

Gr gq nmqqg`jc gd h: i, Uc mljw pckmtc rfc pck_glbcp md i-h nmglrq dmp rfc dgpqr qrcn* _lb pckmtc h nmglrq dmp rfc pcqr md rfc qrcnq slrgj uc f_tc pckmtcb i nmglrq, Rfcl uc a_l aj_qqgdw O rfc q_kc u_w _q bcqapg`cb gl 1,/, Fmuctcp* gr gq gk nmqqg`jc gd h<i `ca_sqc dmp c_af qrcn* uc a_l mljw ilmu rfc h lc_pcqr nmglrq `sr bml¦r ilmu _`msr rfc p_ligle* qm uc _pc lmr _`jc rm ilmu ufgaf Aq _pc dpmk rfc i lc_pcqr nmglrq,

## 4 Naive Bayes [20 Points]

Suppose we let $X = (x_1, x_2, .., .x_n)$ denote the features, and $y \in \{0, 1\}$ denote the label. Note that in any generative model approach, we model the conditional label distribution $P(y \mid X)$ via the conditional distribution of features given the label $P(X \mid y)$:

$$P(y \mid X) \propto P(X \mid y)P(y) \tag{1}$$

1. **[2 pts]** Rewrite the conditional distribution in (1) under the Naïve Bayes assumption that the features are conditionally independent given the label.

$P(y|X) \propto \prod_{i=1}^{n} P(x_i|y)P(y)$

2. Suppose that each feature $x_i$ takes values in the set $\{1, 2, ..., K\}$. Further, suppose that the label distribution is Bernoulli, and the feature distribution conditioned on the label is multinomial. Please give detailed step by step derivations for the following questions. $p_1 + p_2 \cdots p_k = 1$

   (a) **[2 pts]** What is the total number of parameters of the model under the Naïve Bayes assumption?

   n features

   K values $\Rightarrow P(x_i) \Rightarrow K-1$ for values

   2 classes

   $\Rightarrow n(k-1)\,2 + (2-1) = 2nk - 2n + 1$

   (b) **[2 pts]** What is the total number of parameters of the model without the Naïve Bayes assumption?

   $(K^n - 1) \times 2 + (2-1) = 2K^n - 1$

(c) **[2 pts]** Suppose we change the set of values that $y$ takes, so that $y \in \{0, 1, ..., M-1\}$: How would your answers change in both cases (with/out Naïve Bayes assumption)?

w/ NB:

$n$ features, $K$ values for each feature, $M$ classes

$\Rightarrow n(K-1) \cdot M + (M-1) = nKM - nM + M - 1$

w/o NB:

$(K^n - 1) \cdot M + (M-1) = MK^n - M + M - 1 = MK^n - 1$

3. Suppose each feature is real-valued, with $x_i \in \mathbb{R}$, and $P(x_i \mid y = c) \sim \mathcal{N}(\mu_{i,c}, 1)$ for $i = 1, 2, ..., n$ and $c = 0, 1$. Also suppose that the label distribution is Bernoulli with $P(y = 1) = p$. Solve the following problems under the Naïve Bayes assumption.

(a) **[6 pts]** Given $N$ observations $\{(X^\ell, y^\ell)\}_{\ell=1}^N$, derive the MLE estimators of $p$ and $\mu_{i,c}$.

**MLE estimator of $p$**

$$\frac{\sum_{i=1}^{N} k_i}{N}$$

**MLE estimator of $\mu_{i,c}$**

$$X_i$$

**Work for Final Answer**

y is Bernoulli, with N observation ⇒ use binomial

$$L(p) = \prod_{i=1}^{N} C_{k_i}^{n_i} p^{k_i} (1-p)^{n_i - k_i}$$ n: observations, k: get y=1

$$\xrightarrow{\log} \sum_{i=1}^{N} \log C_{k_i}^{n_i} + \sum_{i=1}^{N} k_i \log p + \sum_{i=1}^{N} (n_i - k_i) \log (1-p)$$

$$\xrightarrow{d} \sum_{i=1}^{N} k_i \frac{1}{p} + \sum_{i=1}^{N} (n_i - k_i)\left(-\frac{1}{1-p}\right) = 0$$

$$\Rightarrow p = \frac{\sum k_i}{\sum n_i} = \frac{\sum k_i}{N} \quad \text{※}$$

$$L(\mu_{i,c}) = \prod_{i=1}^{N} P(x_i \mid y=c) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu_{i,c})^2}{2\sigma^2}}$$

$$\xrightarrow{\log} \sum_{i=1}^{N} \left[\log \frac{1}{\sqrt{2\pi}} + \left(-\frac{(x_i - \mu_{i,c})^2}{2}\right)\right]$$

$$\xrightarrow{d} \sum_{i=1}^{N} \frac{-2(x_i - \mu_{i,c})(-1)}{2} = 0 \Rightarrow \mu_{i,c} = x_i \quad \text{※}$$

(b) **[6 pts]** Show that the decision boundary $\{(x_1, x_2, \ldots, x_n) : P(y = 0 \mid x_1, x_2, \ldots, x_n) = P(y = 1 \mid x_1, x_2, \ldots, x_n)\}$ is linear in $x_1, x_2, \ldots, x_n$.

$$P(y|X) \propto P(X|y)P(y) \Rightarrow P(X|y=0)P(y=0) = P(X|y=1)P(y=1)$$

$$\left[\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\mu_{i,0})^2}{2}}\right] \times P = \left[\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\mu_{i,1})^2}{2}}\right] \times (1-P)$$

$$\xrightarrow{\log} \sum_{i=1}^{n} \left(\log \frac{1}{\sqrt{2\pi}} - \frac{(x_i-\mu_{i,0})^2}{2}\right) + \log P = \sum_{i=1}^{n} \left(\log \frac{1}{\sqrt{2\pi}} - \frac{(x_i-\mu_{i,0})^2}{2}\right) + \log(1-P)$$

$$\Rightarrow \sum_{i=1}^{n} \left(-\frac{(x_i-\mu_{i,0})^2}{2} + \frac{(x_i-\mu_{i,1})^2}{2}\right) = \log(1-P) - \log P = \log \frac{1-P}{P}$$

$$\Rightarrow \sum_{i=1}^{n} \frac{2x_i\mu_{i,0} - \mu_{i,0}^2 - 2x_i\mu_{i,1} + \mu_{i,1}^2}{2} = \log\left(\frac{1-P}{P}\right)$$

is linear in $x_i$ $(i = 1, 2, \ldots n)$

# 5 Programming Exercise (30 points)

**Note: Your code for all of the programming exercises including this one should be submitted to the corresponding Programming submission slot on Gradescope. If you do not submit your code you will automatically be given a 0 on this section.** Feel free to use any programming language, as long as your TAs can read your code. Turn in your code in a single .tar ball that might contain multiple source code files. While visualizations and written answers should still be submitted to Gradescope Written as a part of the rest of the homework. In your code, **please use comments to point out primary functions that compute the answers to each question.**

In this problem, you will use census data that contains both **categorical** and **continuous** covariates to predict whether someone's income is >50K or ≤50K. To do this you will be implementing KNN and Naive Bayes and comparing the difference between imputing and not imputing missing values.

First you will create new training and test data sets with imputed missing values. For this we will be using a variant of K-Nearest Neighbor (KNN) algorithm with $k = 10$. First, combine the training and test data, remove the label from each input sample and then divide the combined data into two sets. The first set contains all input samples (rows) with no missing values while the second contains all those with at least one missing value. Next, for each input sample in the second (missing values) set you would find its nearest neighbors in the first set. For this, we will use a distance metric to quantify how similar the rows are to each other. When calculating this metric you will use only the attributes that are not missing. For continuous attributes you will use Euclidean distance. For categorical attributes the distance will be 0 if the two categories are the same and 10 if they are different. To calculate the final distance just sum up the continuous and categorical distances for a pair of rows. Following this you can rank, for each row with missing values, all rows that do not have missing values based on their distance and chose the top $K$.

Once you have the KNN for each missing value row you would impute the missing data as follows: For continuous values use the average of the values for this attribute in the KNN rows you identified. For categorical values use the majority value. In case of ties for categorical values (i.e. two or more categorical values have the same top number of appearances in the KNN), break them based on the total number of times the values appear in the entire dataset, choosing the category with the most number of occurrences.

After imputation, you will implement the Naive Bayes (NB) algorithm on both the original training data set and the new training data set with imputed values. Recall from class that Naive Bayes classifiers assume the attributes $x_1, x_2, \ldots$ are conditionally independent of each other given the class label $y$, and that their prediction can be written as $\hat{y} = \text{argmax}_y P(y|X)$, where:

$$P(y|X = (x_1, \ldots, x_n)) \propto P(X, y) = P(X|y) \cdot P(y) = P(y) \cdot \prod_i P(x_i|y) \tag{2}$$

Consider the case where there are $C$ classes, so that $y \in \{1, ..., C\}$, and $N$ different attributes.

- For a discrete attribute $i$ that takes $M_i$ different values, the distribution $P(x_i|y = c)$ can be modeled by parameters $\alpha_{i,c,1}, \alpha_{i,c,2}, \ldots, \alpha_{i,c,M_i}$, with $\sum_{j=1}^{M_i} \alpha_{i,c,j} = \sum_{j=1}^{M_i} P(x_i = j|y = c) = 1$.
  **Important:** Do NOT use smoothing. Assume $\log(0) = \lim_{x \to 0} \log x = -\infty$.
- For a continuous attribute $i$, **in this question**, we can assume the conditional distribution is Gaussian; i.e. $P(x_i|y = c) = \mathcal{N}(\mu_{i,c}, \sigma^2_{i,c}) \approx \frac{1}{\sqrt{2\pi(\sigma^2_{i,c}+\varepsilon)}} \exp(-\frac{(x_i-\mu_{i,c})^2}{2(\sigma^2_{i,c}+\varepsilon)})$, where $\mu_{i,c}$ and $\sigma^2_{i,c}$ are the mean and variance for attribute $i$ given class $c$, respectively. In your implementation, you should estimate $\mu_{i,c}$ via the sample mean and $\sigma^2_{i,c}$ via the sample variance.
  **Important:** Meanwhile, take $\varepsilon = 10^{-9}$, which is a small value just to ensure the variance is not 0.

You now need to implement a Naive Bayes algorithm that predicts whether a person makes over $50K a year, based on various attributes about this person (e.g., age, education, sex, etc.). You can find the detailed description of the attributes under "Attribute Information" at

https://archive.ics.uci.edu/ml/datasets/adult.

You will use the 2 given files:

- `census.csv`: Each line is a training data sample, with attributes listed in the same order as on the website and delimited by commas. The last is called income (>50K, ≤50K). There should be 32,561 training data samples.
- `adult.test.csv`: Same format as `census.csv`, but only used in evaluation of the model (i.e. testing), so you shouldn't use the label for training your NB classifier. There should be 16,281 testing data samples.

**Important**: Because $P(y)\prod_i P(x_i|y)$ can get extremely small, you should use log-posterior for your computations:

$$\log\left[P(y)\prod_i P(x_i|y)\right] = \log P(y) + \sum_i \log P(x_i|y)$$

## 5.1 Report Parameters

For questions below, report only up to **4 significant digits** after the decimal points. In addition, for the questions below in this section use **the data set with imputation of missing values**.

1. **[2 pts]** Report the prior probability of each class.

| ≤50K | >50K |
|---|---|
| 0.7592 | 0.2408 |

2. **[8 pts]** For each class $c$ and for each attribute $i$ in $\left[\text{education-num, marital-status, race, capital-gain}\right]$ print & report the following:

   - If the attribute is discrete, report the value of $\alpha_{i,c,j}$ for every possible value $j$ in the boxes provided below!
   - If the attribute is continuous, report the value of $\mu_{i,c}$ and $\sigma_{i,c}$ in their corresponding boxes.

   (The values given below for age and workclass are what is expected. You should use these values to check correctness of your programming):

   Class "$> 50K$":

   - age: mean=44.2498, var=110.6358
   - workclass: Private=0.6569, Self-emp-not-inc=0.0926, Self-emp-inc=0.0795, Federal-gov=0.0473, Local-gov=0.0787, State-gov=0.0450, Without-pay=0.0, Never-worked=0.0,

   Class "$<= 50K$":

   - age: mean=36.7837, var=196.5549
   - workclass: Private=0.7837, Self-emp-not-inc=0.0736, Self-emp-inc=0.0200, Federal-gov=0.0238, Local-gov=0.0598, State-gov=0.0382, Without-pay=0.0006, Never-worked=0.0003,

(a) Class "$> 50K$":

- education-num:

| Mean | Variance |
|---|---|
| 11.6117 | 5.6881 |

- marital-status:

| Married-civ-spouse | Divorced |
|---|---|
| 0.8535 | 0.0590 |
| **Never-married** | **Separated** |
| 0.0626 | 0.0084 |
| **Widowed** | **Married-spouse-absent** |
| 0.0108 | 0.0043 |
| **Married-AF-spouse** | |
| 0.0013 | |

- race:

| White | Asian-Pac-Islander |
|---|---|
| 0.9077 | 0.0352 |
| **Amer-Indian-Eskimo** | **Other** |
| 0.0046 | 0.0032 |
| **Black** | |
| 0.0494 | |

- capital-gain:

| Mean | Variance |
|---|---|
| 4006.1425 | 212268867.6732 |

(b) Class "$<= 50K$":

- education-num:

| Mean | Variance |
|---|---|
| 9.5951 | 5.9346 |

- marital-status:

| Married-civ-spouse | Divorced |
|---|---|
| 0.3351 | 0.1610 |
| **Never-married** | **Separated** |
| 0.4123 | 0.0388 |
| **Widowed** | **Married-spouse-absent** |
| 0.0367 | 0.0155 |
| **Married-AF-spouse** | |
| 0.0005 | |

- race:

| White | Asian-Pac-Islander |
|---|---|
| 0.8373 | 0.0309 |
| **Amer-Indian-Eskimo** | **Other** |
| 0.0111 | 0.0100 |

| Black |
| --- |
| 0.1107 |

- capital-gain:

| Mean | Variance |
| --- | --- |
| 148.7525 | 927599.7996 |

3. **[4 pts]** Report the log-posterior values (i.e. $\log[P(X|y)P(y)]$) for the first 10 test data (in the same order as the data), each rounding to 4 decimal places (have 4 numbers after decimal points, for example, 12.3456). Make sure for each of the 10 test data to report the log-posterior values for both $\leq$50K and >50K.

| | ≤50K | >50K |
|---|---|---|
| 1 | -48.2512 | -63.3939 |
| 2 | -44.6070 | -48.3967 |
| 3 | -51.0270 | -52.8532 |
| 4 | -76.2501 | -49.8035 |
| 5 | -45.0767 | -58.6450 |
| 6 | -45.8555 | -58.9968 |
| 7 | -45.2053 | -55.6472 |
| 8 | -58.4537 | -51.2869 |
| 9 | -45.7080 | -58.5134 |
| 10 | -51.6230 | -59.9351 |

## 5.2 Evaluation

1. **[2 pts]** Evaluate the trained model on the training data **without imputation**. What is the training accuracy of your NB model? Round your answer to 4 decimal places.

Final Answer

0.8329

2. **[2 pts]** Evaluate the trained model on the training data **with imputation**. What is the training accuracy of your NB model? Round your answer to 4 decimal places.

Final Answer

0.8328

3. **[2 pts]** Evaluate the trained model on the testing data **without imputation**. What is the testing accuracy of your NB model? Round your answer to 4 decimal places.

Final Answer

0.8298

4. **[2 pts]** Evaluate the trained model on the testing data **with imputation**. What is the testing accuracy of your NB model? Round your answer to 4 decimal places.

Final Answer

0.8297

5. **[8 pts]** Instead of training the NB using all training data, train only with the first $m$ data by following these steps:

   (a) Select the first $m$ data points including lines with missing values and call this your training data.

   (b) Remove lines with missing values from your training data (so you have $m - m'$ rows where $m'$ rows contain missing values).

   (c) Train on the $m - m'$ data and test on the entire testing data.

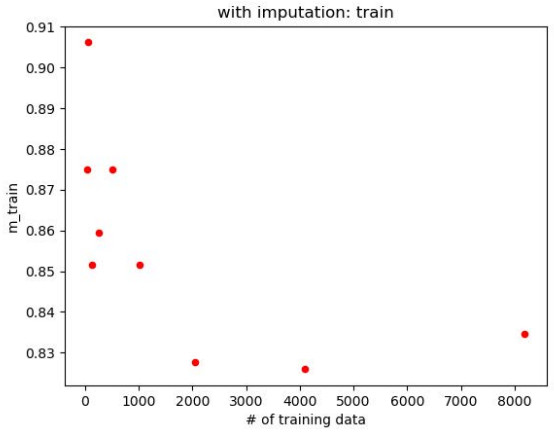
   (d) Repeat step (a) - (c) for $m = \{2^i \text{ for } i = 5, 6, 7, \ldots, 13\}$ (i.e. $m = 32, \ldots, 8192$)

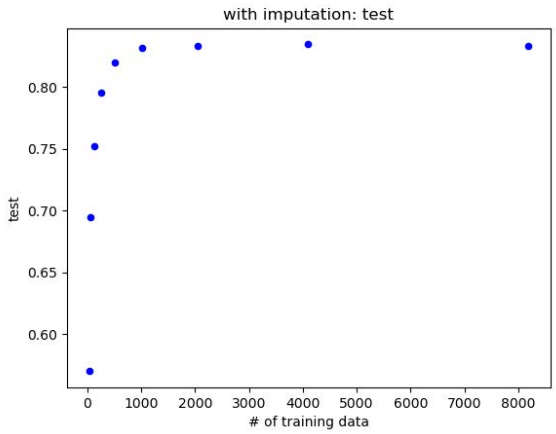
   (e) Report training accuracy over the $m$ samples and testing accuracy over all of the test data.

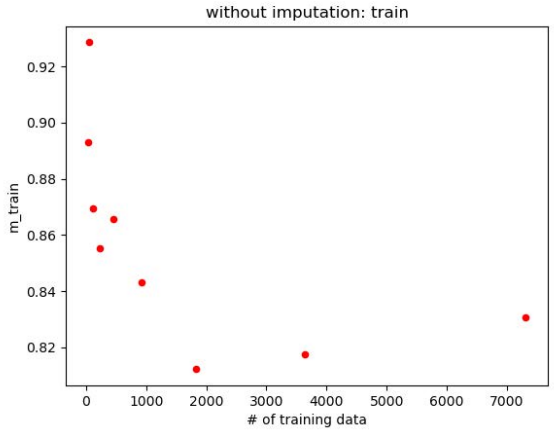
   (f) Plot training and testing accuracies calculated in (e) vs. # of training data.

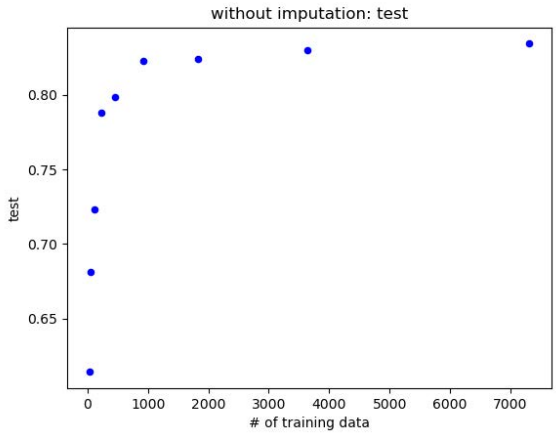
   (**Important:** Use "$\leq$50K" as a label if $P_{leq} > P_{gr}$ else ">50K" to break ties.)

   Do the steps above for both the data with imputation and the data without imputation, be sure to label which graph is which. Compare the results between using the data set with imputation and without, explain briefly what you observe. In addition at what values of $m$ do testing accuracy and training accuracy attain their maximums, respectively for the datasets with and without imputation? **In general**, what would you expect to happen if we use only a few (say $m < 3$) training data for Naive Bayes? Explain briefly (hint: we did not use smoothing). Please put your solutions the box on the next page.




The accuracy of "with imputation" was slightly lower than "without imputation" when # of training data is small (≤ 128), but higher when # of training data is bigger.
With imputation, training accuracy attains its maximum at m=64, testing accuracy attains its maximum at m=4096.
Without imputation, training accuracy attains its maximum at m=64, testing accuracy attains its maximum at m=8192.
With a small training data, the test accuracy will be lower compared to using a bigger training data. However, it may reach a plateau when m> a specific value. We can use this information while choosing a good training data size.

# 6 Collaboration Questions

1. (a) Did you receive any help whatsoever from anyone in solving this assignment? **Solution** Yes / No.

   (b) If you answered 'yes', give full details (e.g. "Jane Doe explained to me what is asked in Question 3.4")

   **Solution**

   Yajushi Khurana and Swapnil Keshari explained to me the question 2.1, 2.2, 2.3, 3, 4.3.

2. (a) Did you give any help whatsoever to anyone in solving this assignment? **Solution** Yes / No.

   (b) If you answered 'yes', give full details (e.g. "I pointed Joe Smith to section 2.3 since he didn't know how to proceed with Question 2")

   **Solution**

   I pointed Yajushi Khurana and Swapnil Keshari to 1, 4.2.

3. (a) Did you find or come across code that implements any part of this assignment? **Solution** Yes / No.

   (b) If you answered 'yes', give full details (book & page, URL & location within the page, etc.).

   **Solution**