EECE 5644
Aleksei Krotov

Homework 2.

The code used for this homework, as well as for the other course homeworks, are in this repository.

Q1. Using the likelihood of a single observation $p(D|\theta) = p(x_k|\theta)$ and assuming $X \sim N(\mu, \sigma^2)$, the ML estimates of the mean and variance are $\hat{\mu} = x_1$ and $\sigma = \frac{1}{1} \cdot (x_1 - \mu) = 0$, while the unbiased estimate of variance is $[\frac{0}{0}]$. The biased (ML) estimate of variance corresponds to an empty, non-existent, or infinitely small entity, while the unbiased estimate of variance corresponds to a non-existent and uncertain entity. Neither is strictly wrong, but the unbiased one makes more sense not adding the connotation of a negligibly small amount.

Q2. For a multinomial distribution with six categories and a dataset $D = \{x_1, x_2, ..., x_N\}$, assumind i.i.d sampling, likelihood of such dataset is $p(D|\theta) = p_{11} \cdot p_{12} \cdot ... \cdot p_{1N_1} \cdot p_{21} \cdot p_{22} \cdot ... \cdot p_{kN_k} = \prod_{k=1}^{K} \prod_{i=1}^{N} p_{ki} = \prod_{k=1}^{K} p_k^{\sum_{i=1}^{N} \mathbf{1}_{x_i=k}} = \prod_{k=1}^{K} \theta_k^{N_k}$.

If a prior probability is assumed to follow Dirichlet distribution $p(\theta) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$, the posterior probability shall be $p(\theta|D) \propto p(D|\theta)p(\theta) \propto \prod_{k=1}^{K} \theta_k^{N_k + \alpha_k - 1}$, i.e. following Dirichlet distribution with $\alpha'_k = \alpha_k + N_k$.

For $\alpha_1 = \alpha_2 = ... = \alpha_K = 1$, the posterior probability shall be $\prod_{k=1}^{K} \theta_k^{N_k}$. Assuming that the *k*-faceted die is unbiased, $\theta_1 = \theta_2 = ... = \theta_k = \frac{1}{k}$, therefore the posterior probability is not affected by the prior, but only by the data, and the prior itself follows a distribution $p(\theta) \propto \prod_k 1 \propto 1$, i.e. a uniform distribution.

If only one sample $\{x_1\}$ is available and the prior is uninformative, the estimates become:

$$\hat{\theta}_k^{ML} = \begin{cases} 1, & k = x_1 \\ 0, & otherwise \end{cases}, \quad \hat{\theta}_k^{MAP} = \begin{cases} \frac{1+1-1}{1+6-6} = 1, & k = x_1 \\ \frac{0+1-1}{1+6-6} = 0, & otherwise \end{cases}, \quad \text{and}$$

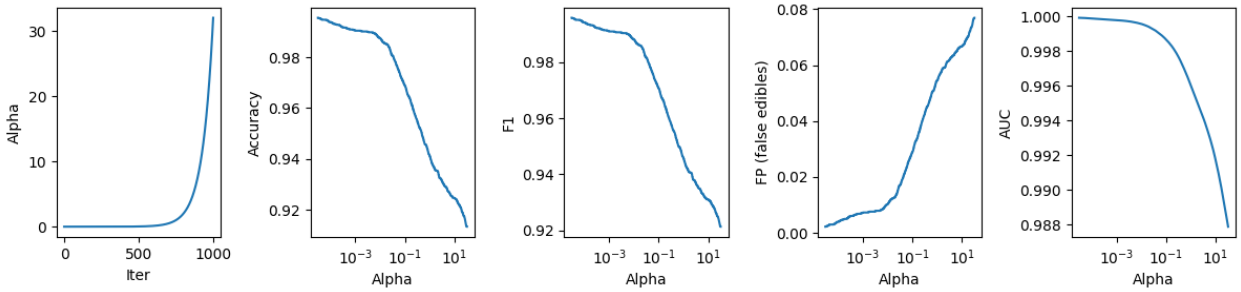
$$\mathrm{E}[\theta_k | D] = \begin{cases} \frac{1+1}{1+6} = \frac{2}{7}, & k = x_1 \\ \frac{0+1}{1+6} = \frac{1}{7}, & otherwise \end{cases}.$$ The MAP estimate approaches the ML estimate under an

uninformative prior and both cannot capture the probability of an unseen category, while the mean of a posterior probability avoids that issue and makes more sense.

Q3.

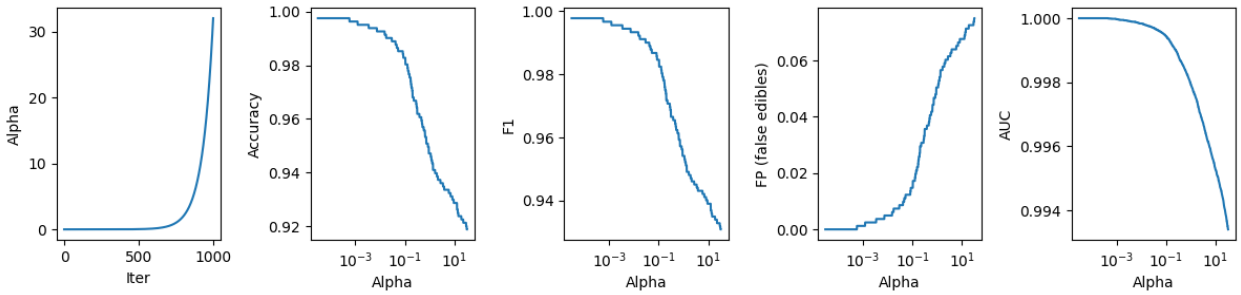
The decision rule, using Naïve Bayesian classifier with a Dirichlet prior is maximizing the mean of the posterior probability conditioned on class and can be expressed as follows: $c = \arg\max_c \left[p(y = c \mid \mathbf{x}, D \right] = \arg\max_c \left[\bar{\pi}_c \cdot \prod_{j=1}^{D} \theta_{jc}^{\mathbf{1}_{x_j=1}} (1-\theta_{jc})^{\mathbf{1}_{x_j=0}} \right]$, where $\pi_c = \frac{N_c + \alpha_c}{N + \alpha_0}$, N_c and N are numbers of the category seen in the dataset, conditioned on class, and number of observations respectively, α_c and α are category value and all categories from the Dirichlet distribution, aka pseudo-counts, and θ_{jc} is the empirical probability of seeing a category, class conditioned, in the dataset.

The mushrooms dataset from the UCI repository was used, where character categories were converted to numerical ones via provided script *process_mushrooms*.
First a train-test split 80/20 was used to train and test the Naïve Bayes Categorical classifier, using *CategoricalNB* from *scikit* library. To account for category values that might be missing in the training data, minimum number of categories was set to the number of categories for each feature respectively, from the original dataset. Laplace parameter α was varied from 2^{-15} to 2^{5} using 1000 values on a logarithmic scale.
The classifier performance at different α values are shown below.



The area under the ROC curve (the rightmost panel) is maximized at the minimum $\alpha = 2^{-15}$, and so are accuracy and F1-score, while false-positives (essential for this particular classification) are minimized at the same value at 0.22%

Splitting test/train as 10/90 resulted in overall higher accuracy scores and the same trend of classifier parameters monotonically deteriorating with increasing α, that maintained both at split randomizer seeds 42, 20, and 10. At $\alpha = 2^{-15}$, accuracy was 99.75%, F1 score 99.78%, AUC 100.0%, and 0.0% of false-positives (poisonous mushrooms classified as edible).

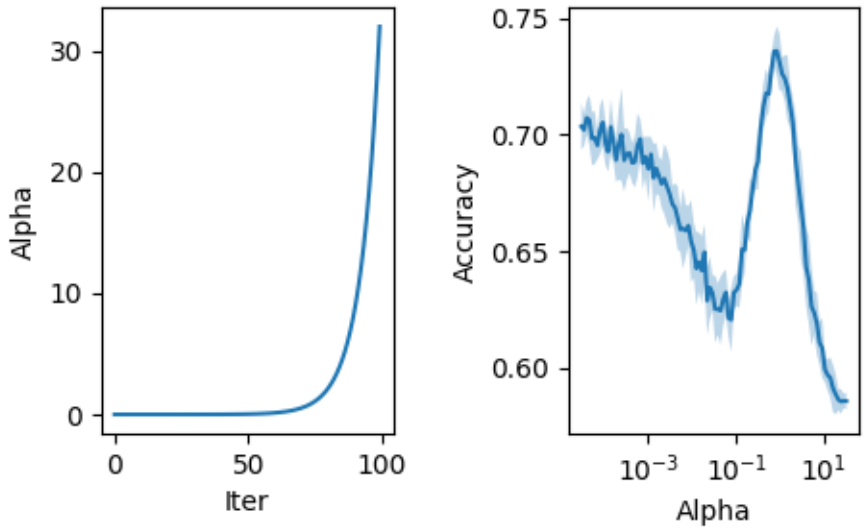


While the dataset is balanced with respect to classes (3916 poisonous/unknown samples and 4208 edible samples), the false-positive classifications (poisonous identified as edible) appears more important than false-negatives (edible identified as poisonous). Therefore, for this dataset, AUC, as well as additionally computed false-positive ratio provide better insight at the classifier performance, than accuracy.

Q4. For this problem, multiple labeled articles from dataset Sentence Classification were used, pre-processed into "bag of words" representation using provided script *process_sentences*. The data consisted of five classes (MISC, AIMX, OWNX, CONT, and BASE), 4164 words (features) and 3117 sentences (observations).

The data were split into train/test sets as 80/20 and a multinomial Naïve Bayes classifier was trained on the training set, using varying smoothing parameter between 2^{-15} and 2^{5} in the logarithmic scale. The test set was then predicted and classifier accuracy was evaluated. After repeating the split 10 times at different randomizer seed, the average relation of accuracy and smoothing parameter was obtained (see the plot below, showing the mean accuracy and one standard deviation as a shaded band). Maximum accuracy 0.731±0.011 was reached at $\alpha = 0.9656$

.



The five words (features) with the highest posterior probability at smoothing parameter maximizing accuracy were obtained using vocabulary from *process_sentences*. These words (in the order of reducing probability) are given below:

Class					
MISC	citation	of	to	and	the

AIMX	citation	of	to	a	the
OWNX	of	to	we	and	the
CONT	to	and	a	in	the
BASE	symbol	of	to	a	the

The abundance of stop words and prepositions, as well as their overlap, might undermine the classifier's accuracy. A better performance can be potentially reached, by excluding the stop-words or reducing their weight, and by inspecting the relation of each word with the class label, via for example, mutual information or chi-squared criterion and their pre-weighting based on predictive power.