Machine Learning - Homework 6 Report

Lukas Johannes Ruettgers (2023403372)

November 8, 2023

1 Objective

The experiment's objective is firstly to examine the performance of a Naïve Bayes Classifier (NB) on a multi-feature dataset. Secondly, we investigate the influence of a Minimal Risk Decision on the final classification results. The dataset originates from https://www.kaggle.com/c/widsdatathon2020/data and contains medical information of patients at the Intensive Care Unit (ICU) of US-American hospitals. The patients x_i were classified into two groups ω_i depending on their survival s_i during their stay at the ICU.

The model is supposed to estimate the probability of survival ($s_i = 1$) or death ($s_i = 0$) respectively for any patient x_i represented by the feature schema of the given dataset.

2 Experiment setup

2.1 Model

In contrast to the earlier models, the Naïve Bayes Classifier belongs to the probabilistic classifiers which assume that the samples x of each class ω_i originate from a probability distribution $p(x \mid \omega_i)$. This distribution is commonly referred to as the *conditional density* of x given ω_i. Of course, $p(x \mid \omega_i)$ is not known in the beginning. Since x comprises multiple features, the size of the hypothesis space for such a conditional density $p(x \mid \omega_i) = p(x_1, \dots x_{n_{FT}} \mid \omega_i)$ is exponential in n_{FT}. If not given more knowledge about the features, the exhaustive search to find even a nearly optimal hypothesis is mostly infeasible. To approximate such a good hypothesis, Naïve Bayes assumes pairwise conditional independence of the features, namely

$$p(x_j \mid x_1, \dots, x_{n_{FT}}, \omega_i)) = p(x_j \mid \omega_i) \text{ for all } 1 \leq j \leq n_{FT}.$$

If this assumption would hold, each feature would be independently distributed and would therefore follow a univariate probability distribution. Even though this is rarely true in reality, it reduces the problem's complexity to finding n_{FT} *univariate* conditional densities. For each feature, a different (parametrized) probability density model may now be optimized by a *maximum likelihood* estimation or other methods. Having estimated such densities, Bayes' rule paves the way to express the *posterior* probability that an observed sample x belongs to a class ω_i as

$$P(\omega_i \mid x) = \frac{\left(\prod_{j=1}^{n_{FT}} P(x_j \mid \omega_i)\right) \cdot P(\omega_i)}{P(x)}.$$

By comparing these posterior probabilities for different classes ω_i, we can follow Bayes' decision rule and pick the class for which the posterior probability is the highest. To that end, we nevertheless require the prior probability $P(\omega_j)$ for each class, which needs to be

estimated by domain knowledge. If such domain knowledge is not given, the priors can also be considered as a variables and estimated with Maximum a Posteriori (MAP) or towards their respective worst-case values. Since $P(x)$ is invariant to ω_i for a fixed x, this probability can be ignored in the comparison of the posteriors.

2.2 Preprocessing

The training dataset and the testing dataset comprises $n_{TrSamp} = 5000$ and $n_{TsSamp} = 1097$ samples respectively. Each sample is described by $n_{FT} = 108$ features subject to binary, numeric and integer domains. While the binary values were solely converted into integers, the values of numeric and integer features were all normalized to $[0, 1]$, since their values varied largely in size and units. This was achieved by linearly rescaling each feature value x to $\frac{x-x_{min}}{x_{max}-x_{min}}$, where x_{min} and x_{max} are the maximum and minimum values of the respective feature.

2.3 Cross Validation

Besides training the model with the entire training dataset, the bias of the models was assessed via 10-fold cross validation (CV). The training data set was split into ten equally sized intervals and the 10-fold CV error was computed as the mean accuracy of the predictions of the models trained on their respective testing interval.

2.4 Implementation details

In this experiment, the Naive Bayes classifier models of the `scikit-learn` package are used. Unfortunately, these models are not able to estimate different probability distribution families for different features. All features are either assumed to follow a Gaussian distribution (GaussianNB), a multinomial distribution (MultinomialNB) or a Bernoulli distribution (BernoulliNB). While a Gaussian distribution is a suitable choice for continuous real valued variables such as numeric values, the Multinomial and the Bernoulli distribution are intended to model discrete and binary features respectively.

Since the given dataset comprises both binary and numeric features, I decided to partition the feature matrix into numeric and binary features. First, a Gaussian Naïve Bayes Classifier (GaussianNB) was solely trained on the numeric features, whose fine domain allows precise quantification of scores and metrics. Secondly, both a Multinomial and Bernoulli Naïve Bayes Classifier were tried out on the binary features, which mostly express the membership of patients in certain diagnosis groups. Then, the prediction probabilities of both models serve as a two-feature, real-valued input data for a third model, called *composite model* in the following. This composite model is a GaussianNB and hence fits the given probabilities to a Gaussian distribution to make the final decision. To compare the composite models' performance each given one of the MultinomialNB and BernoulliNB as submodel for the binary features, we consider the training, validation and testing accuracy, which is defined as $\frac{TP+TN}{TP+FP+FN+TN}$.

For the reason that 48.99% of the samples in the training dataset and 49.19% of the samples in the test dataset died during their stay in the ICU and therefore belong to class ω_0, we find a prior probability of 50% for each class to to be sufficiently close to the true unknown prior. However, we also left the parameter `fit_prior` in the MultinomialNB and BernoulliNB activated to enable the model to adjust the prior if needed.

2.5 Minimal Risk Decision

The Naïve Bayes classifier maximizes the accuracy and in particular weighs false positive (FP) and false negative (FN) errors equally. In reality, different decision outcomes mostly

entail different risks or costs. For that reason, we investigate how the predictions change if we assign more realistic risks to each possible of the four decision scenarios TP, FP, FN and TN. In the following, we denote each of these scenario risks by λ_{ij}, where i represents the class which the model predicted x to belong to while j represents its true class.

The objective now shifted to make the decision i which minimizes the term

$$P(\omega_0 \mid x) \cdot \lambda_{i0} + P(\omega_1 \mid x) \cdot \lambda_{i1}.$$

Rearranging these two terms for the two possible decisions, we obtain the following decision rule:

$$\text{If } \frac{P(\omega_1)}{P(\omega_0)} \cdot \frac{\lambda_{01} - \lambda_{11}}{\lambda_{10} - \lambda_{11}} \lessgtr \frac{P(\omega_0 \mid x)}{P(\omega_1 \mid x)}, \text{ then choose } \begin{matrix}\omega_0 \\ \omega_1\end{matrix}.$$

Regarding the critical context in which the model might be used, its predictions may indeed entail consequences that reach as far as giving up on a patient's life. Many ICUs are overcharged with patients and must carefully schedule where to invest their energy and resources into. For the following arguments, we assume that the objective of ICUs is to save as many lives as possible, independent of the patient's identity. We hence assume as well that the life of each patient at a given ICU has an equal weight. Weighing the value of life of a human is indeed a hard problem. But assuming an equal weight of every life, it simplifies the comparison between different lifes. In a very abstract notion, a human contributes to the world by investing energy (kinetic, emotional, mental...) into the problems of our world during his lifetime. This notion will guideline us in the upcoming arguments. Over and above, we must assume that the model is the only decision-maker for the survival chance of a patient.

	ω_0	ω_1
α_0	**0.05**	10
α_1	1	0

(a) Option 1

	ω_0	ω_1
α_0	**0.2**	10
α_1	1	0

(b) Option 2

Figure 1: Risk values λ_{ij} for the classification decision α_i on a sample x which truly belongs to class ω_j ($0\widehat{=}$ died, $1\widehat{=}$ survival).

To normalize the risks, let $\lambda_{10} = 1$ be the cost incurred in the scenario where the model predicts that a patient x survives if he (or she) is further medically treated at the ICU, but nonetheless the patient dies. This means that resources and staff have been invested into this patient without any profit. These resources could have been invested into the treatment of other patients instead and therefore this decision causes opportunistic costs. If the patient survives and can get out of the hospital, he is able to invest energy into the world. Let us assume that these contributions generate a long-term surplus which exceeds the costs of the entire ICU treatment of the patient from the time the model makes its prediction. While one could also consider negative costs, we will only assume non-negative costs and consequently set the risk $\lambda_{11} = 0$.

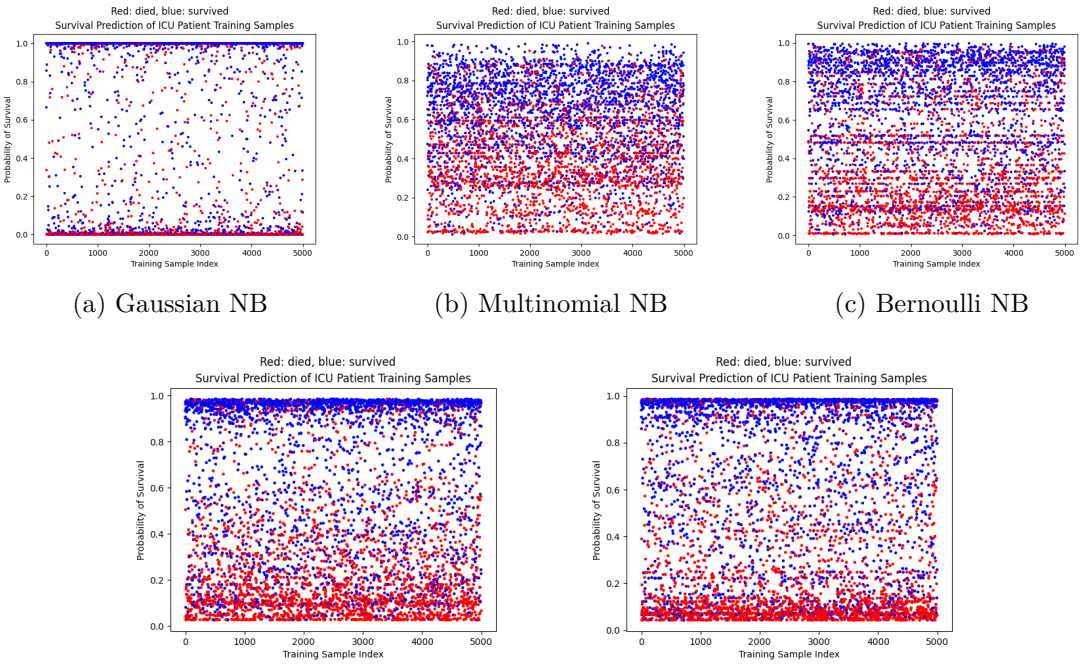
On the contrary, consider the case where a patient would have survived but is let down by the ICU staff because the model predicts him to die. This will not only cause opportunistic costs regarding the potential contributions this patient could yet have made to the world, but his family and friends might also incur strong emotional pain which consumes energy they could have invested into the world. We therefore set the cost to be $\lambda_{01} = 10$. Of course, the family members and friends do not surely know that the patient would have survived, but even a small hope suffices to cause grief. Consequently, this grief will also occur in the case where the patient indeed dies. However, this grief may in this case arise less from the decision of the model but rather from the general fact that the

patient passed away, since the evidence for the poor health status of the patient might also be more compelling. We try out both $\lambda_{00} = 0.05$ and $\lambda_{00} = 0.2$ to gauge the impact on the change of this risk value. All in all, Figure 1 depicts both risk value assignments.

3 Results

3.1 Training and Validation

How the assumption of conditional independence already hinders the NB model to detect correlations between features to better fit the data is already indicated by the results in the training process. Because of the independent consideration of each feature distribution, the Gaussian NB model finds its optimal decision rule to be strongly biased to predicting death over survival. While only falsely predicting survival for 399 patients that truly died, it falsely classifies 902 patients in the opposite direction, which in total yields an accuracy of 73.98%. Fig. 2a illustrates that the learned individual distributions of the numeric features in average exhibit a stronger variance for patients that have survived than for patients that died. On the other side, both the Multinomial (Fig. 2b) and the Bernoulli (Fig. 2c) classifier struggle with distinguishing the patients based solely on the binary features. This suggests that the individual distributions of the binary features conditioned on the two classes are in average too close to tell the samples apart. While these models only achieve an accuracy of 69.78% and 70.88% respectively, their predictions are not as biased as the ones of the Gaussian NB. In fact, the Multinomial NB's predictions comprise 676 False Positives and 780 False Negatives while the Bernoulli NB even seems to exhibit a small bias towards the class of survived patients, having 810 FPs and 701 FNs respectively. Even though the Multinomial performs slightly worse than the Bernoulli



(a) Gaussian NB (b) Multinomial NB (c) Bernoulli NB

(d) Composite NB with Multinomial submodel (e) Composite NB with Bernoulli submodel

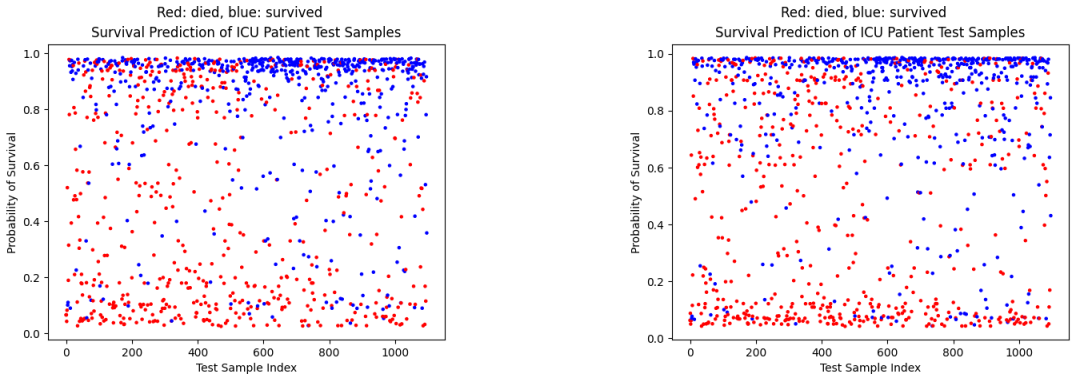
Figure 2: The models' prediction probabilities on the training datasets. Red dots indicate patients that died, blue dots indicate patients that survived. The first line comprises the three submodels and the second line the composite models.

NB and even seems to exhibit a similar bias as the Gaussian NB, the composite model with the Multinomial NB as submodel (Fig. 2d) catches up on its counterpart and even achieves a slightly better accuracy of 75.92% in comparison to the composite model with the BernoulliNB as submodel (Fig. 2e), whose accuracy amounts to 75.84%. Also in cross validation, the accuracy of the composite model with the multinomial submodel slightly but insignificantly surmounts the accuracy of the composite model with the Bernoulli submodel. Their accuracies amount to 75.82% and 75.60% respectively.

The improvement of the composite model suggests that both the numeric features and the binary features cover some disjoint meaningfulness for the patients survival chances.

3.2 Testing

Even though the composite models rectifies the bias of the Gaussian submodel and align the False Positives and Negatives, it fails to maintain such a balanced classification for the test dataset. Here, for both the multinomial and the Bernoulli submodel, the bias



(a) NB with Multinomial submodel

(b) NB with Bernoulli submodel

Figure 3: The composite models' prediction probabilities on the test datasets. Red dots indicate patients that died, blue dots indicate patients that survived.

shifts towards predicting survival, so that only 69 and 66 respectively out of 1097 patients are falsely predicted to die (Fig. 4). The prediction probabilities in Fig. 3 depict this bias, as the blue points, which represent the survived patients, mostly keep a high survival prediction, while the red points are spread across the entire probability range. In total, the accuracy drops to around 70% in both models.

3.3 Minimal Risk Decision

If we further extend our decision rule to consider the risk values in Fig. 1, the bias towards predicting survival increases even more. Mathematically speaking, the risk values raise the required threshold of the predicted probability that a patient dies in order for him to be classified into ω_0. While this threshold amounted to 50% in the normal Bayesian decision rule where both errors weigh equally, the threshold now becomes $t_1 = \frac{10-0}{1-0.05} \approx 10.53$ for the decision table in Fig. 1a and $t_2 = \frac{10-0}{1-0.2} = 12.5$ for the decision table in Fig. 1b. Since the priors for the two classes were assumed to be 50%, they cancel out in the upper terms. In order to decide that a patient will die, the models now require to predict death with a probability of $\frac{t_1}{1+t_1} \geq 91.32\%$ and $\frac{t_2}{1+t_2} \geq 92.59\%$ respectively. This shrinks the number of test samples predicted to die down to 7 out of 1097 for the composite model with the multinomial submodel and $\lambda_{00} = 0.2$ (Fig. 4). Here, the Bernoulli and Multinomial models exhibit more disparities: the composite model with the Bernoulli submodel more oftenly predicts death on the test samples. Obviously, the accuracy for both models diminishes

to around 58%. The higher the risk of predicting death on patients, the more certain the model must be to decide for this class and the more does the accuracy decrease towards 50%.

Model Parameters		**Prediction Results**				
Binary NBC	λ_{00}	TP	FP	FN	TN	Accuracy
Bernoulli	-	**481**	268	**66**	282	0.6955
Multinomial	-	478	**251**	69	**299**	**0.7083**
Bernoulli	0.05	524	417	23	133	0.5989
Bernoulli	0.2	530	439	17	111	0.5843
Multinomial	0.05	539	454	8	96	0.5789
Multinomial	0.2	540	463	7	87	0.5716

Figure 4: Test results of the investigated models. Models are categorized by their kernel function (Kernel), regularization rate (RR) and their kernel coefficient γ.

Our models are trained to make decisions based on the data. There more we take into account other hidden values such as ethical or political considerations, the less important the data becomes for the decision. Especially in such a sensitive context where ethical aspects play an important role, the experiment results show that the model's predictions should only be taken into account when the predictions are made with a high certainty. However, when assuming conditional independence on the features as the Naïve Bayes does, such a certainty will be seldom reached. For such scenarios, the computational cost of stronger models should surely be worth it.