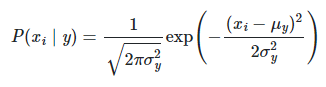
**1.b**

**Gaussian Naive Bayes**

This approach is built on the assumption of a normal distribution of probabilities. **GaussianNB** implements the Gaussian Naive Bayes algorithm for classification.

For example: It means, that spam and not-spam classes of messages have frequencies of the words from vocabulary distributed by the Gaussian law:

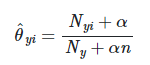


The parameters σy and μy are estimated using maximum likelihood.

**Multinomial Naive Bayes**

Multinomial classification suits best for the discrete values like word counts. **MultinomialNB** implements the naive Bayes algorithm for multi-nomially distributed data, and is one of the two classic naive Bayes variants used in text classification. The distribution is parametrized by vectors θy=(θy1,…,θyn) for each class y, where n is the number of features (in text classification, the size of the vocabulary) and θyi is the probability P(xi∣y) of feature i appearing in a sample belonging to class y.

The parameters θy is estimated by a smoothed version of maximum likelihood, i.e. relative frequency counting:



where Nyi=∑x∈Txi is the number of times feature i appears in a sample of class y in the training set T, and Ny=∑i=1n Nyi is the total count of all features for class y.

Setting α=1 is called Laplace smoothing, while α<1 is called Lidstone smoothing.

**Bernoulli Naive Bayes**

**BernoulliNB** implements the naive Bayes training and classification algorithms for data that is distributed according to multivariate Bernoulli distributions; i.e., there may be multiple features but each one is assumed to be a binary-valued (Bernoulli, boolean) variable. Therefore, this class requires samples to be represented as binary-valued feature vectors; if handed any other kind of data, a BernoulliNB instance may binarize its input (depending on the binarize parameter). The decision rule for Bernoulli naive Bayes is based on



Bernoulli formula is close to the multinomial one, though the input is the set of binary values (eg: the word is present in the message or not) instead of the set of frequencies. So, the algorithm explicitly penalizes the non-occurrence of a feature (word in the message is absent in the vocabulary) while the multinomial approach uses the smoothing parameter for the absent values.