Наивный Байесовский классификатор (Naïve Bayes Classifier, NBС) является одним из примеров использования методов векторного анализа. Данная модель классификации базируется на понятии условной вероятности принадлежности документа  классу .

NBС – один из самых часто используемых классификаторов, из-за сравнительной простоты в имплементации и тестировании. В то же время, наивный Байесовский классификатор демонстрирует не худшие результаты, по сравнению с другими, более сложными классификаторами.

В основе наивного Байесовского классификатора лежит теорема (или формула) Байеса:

Для данной модели, документ – это вектор: , где - вес -ого термина, а – размер словаря выборки. Таким образом, согласно теореме Байеса, вероятность класса для документа будет:

Таким образом, вычисляется условная вероятность для всех классов.

Наиболее вероятный класс , которому принадлежит документ  тот, при котором условная вероятность принадлежности документа классу максимальна:

По теореме Байеса:

и, так как

то

Знаменатель может быть опущен, так как для одного и того же документа вероятность будет одинаковой, а это значит, что ее можно не учитывать.

Для наивного Байесовского классификатора определено существенное допущение – предполагается, что все признаки  документа  независимы друг от друга. Из-за этого допущения модель и получила название «наивная». Это очень серьезное упрощающее допущение и, в общем случае, оно неверно, но наивная Байесовская модель демонстрирует неплохие результаты, несмотря на это. Предполагается так же, что позиция термина в предложении не важна. Как следствие, условную вероятность

для признаков , можно представить, как

Таким образом, для нахождения наиболее вероятного класса для документа с помощью наивного Байесовского классификатора, необходимо посчитать условные вероятности принадлежности документа для каждого из представленных классов отдельно и выбрать класс, имеющий максимальную вероятность:

Теперь необходимо оценить  и . Оценить вероятность класса несложно: является отношением количества документов класса в обучающей выборке к общему количеству документов в выборке.

где  определяется как отношение количества терминов  в классе общему количеству терминов в этом классе. – словарь обучающей выборки.

Существует небольшая проблема, связанная с этой формулой. Если в тестовом наборе встретится слово, которое не встречается в наборе обучающих документов, то вероятность этого слова для любого из классов будет равна нулю. Поскольку

то и вероятность принадлежности документа любому из классов также будет равна нулю, что, конечно, неправильно. Для решения этой проблемы обычно используют так называемое аддитивное сглаживание (add-1 smoothing или сглаживание Лапласа). Идея add-1 сглаживания заключается в том, что к частотам появления всех терминов из словаря искусственно добавляется единица. Получается, что термины, которые не присутствовали в документах обучающей выборки, получают незначительную, но не нулевую вероятность появления и, тем самым, дают возможность определить документ в какой-либо из классов.

Здесь - количество слов в словаре обучающей выборки.

Как было сказано выше, документ представляется в виде вектора для классификации: , где   - веса для каждого из терминов словаря выборки. может быть количеством вхождений термина в документ , или же может быть задано бинарно. Для бинарного вектора число вхождений термина не имеет значения, важен лишь факт появления в документе .

Можно заметить, что для относительно больших текстов, вероятность представляет собой произведение большого количества очень маленьких дробей. Для того чтобы избежать потери точности, можно заменить произведение вероятностей суммой логарифмов вероятностей.