Кроме этого, в метрических алгоритмах часто используются так называемые меры близости. В отличие от метрик, которые тем меньше, чем объекты более похожи, меры близости увеличиваются при увеличении похожести (близости) объектов.

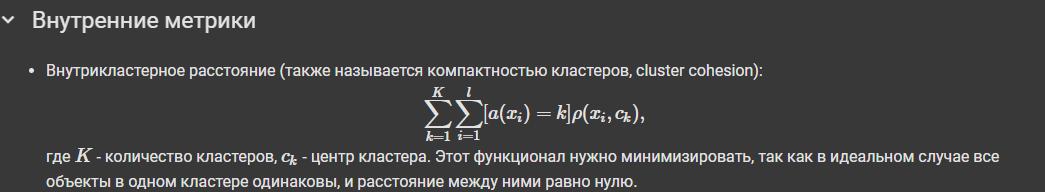
Получается, что при работе с алгоритмом kNN исследователю требуется подобрать три параметра - количество соседей k, метрика расстояния и способ вычисления весов.

Рассматриваемый метод, несмотря на положительные стороны в виде легкости интерпретации, простоты и удобства использования, обладает некоторыми минусами, в частности, он плохо работает на датасетах с большим количеством признаков.

Например, если мы имеем три объекта, при этом второй отличается от первого только значением одного признака, но значительно, а третий отличается от первого незначительно в каждом признаке, расстояния от первого объекта до второго и третьего могут совпадать. Несущественные различия в каждом признаке могут иметь большее значение, чем большое различие в одном признаке. Такое поведение в ряде случаев будет нежелательным.

Второй пример - случай, когда количество признаком сравнимо с количеством объектов. В этом случае может возникнуть ситуация, когда расстояния между любыми двумя объектами почти одинаковы. В двумерном пространстве (на плоскости) три точки могут располагаться по вершинам равностороннего треугольника, при этом расстояния между ними будут равны; в трехмерном пространстве то же самое справедливо для четырех точек на вершинах тетраэдра - расстояние между любыми двумя точками будет одинаково. В общем случае это означает, что в n-мерном пространстве можно выбрать n+1 точку так, чтобы расстояние между любыми двумя точками было одинаковым.

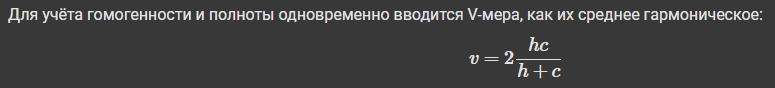
Третий пример - так называемое "проклятие размерности". Суть его заключается в том, что при наличии n бинарных признаков в пространстве признаков будет возможно 2n различных объектов вида x=(0,1,0,0,1,...,1), и размер обучающей выборки, необходимый, чтобы описать все пространство объектов (то есть все возможные комбинации таких признаков) также будет порядка 2n. Чтобы покрыть не все пространство, а долю объектов α (то есть долю объема α) нужно будет описать гиперкуб с длиной ребра ep=αn. Например, в 10-мерном пространстве признаков чтобы покрыть 1% объема нужно взять гиперкуб с длиной e10(0,01)=0.63, то есть взять окрестность длиной больше половины ребра исходного пространства. Чем больше признаков, тем меньше будет область, которая покрывается во время поиска на заданном расстоянии. Таким образом, при сохранении требований по точности нахождения объекта в пространстве, количество требуемых данных для этого при увеличении количества признаков растет экспоненциально.



Выше – сумма расстояний всех объектов до центров их кластеров

Гомогенность измеряет, насколько каждый кластер состоит из объектов одного класса.

Полнота измеряет, насколько объекты одного класса относятся к одному кластеру.



Алгоритмы снижения размерности пространства признаков делятся на две группы - **отбор признаков** (то есть исключение наименее важных признаков) и **понижение размерности** путем формирования новых признаков на основе старых.