Школа анализа данных
Машинное обучение, часть 1
Домашнее задание №3

Кошман Дмитрий

Задача 1 (1.5 балла). Метрики качества.
Матожидание ошибки до костиных преобразований:

$$E = \sum(p_i(1 - y_i) + (1 - p_i)y_i)$$

И после:

$$E' = \sum(p_i I[y_i < 0.5] + (1 - p_i)I[y_i > 0.5])$$

Поскольку я ничего не знаю, о том, как формировались выборка и предсказания, я могу сделать только самые общие выводы. Например, можно сказать, что при сдвигании прогноз в один из концов, Костя повышает уверенность в своих ответах. Соответственно, это улучшит качество, если предсказания были изначально хорошего качества, и ухудшит в противном случае. Продемонстрирую это на примерах. Пусть Костя предсказывает вероятность выпадения орла при подкидывании смещенной монетки с истинной вероятностью 0.55. Пусть данных о бросаниях много, и максимально правдоподобная оценка оказалась равна 0.53. Тогда до преобразования ошибка равна $0.55 * (1 - 0.53) + 0.45 * 0.53 = 0.497$, после - $0.55 * 0 + 0.45 * 1 = 0.45$, качество улучшилось. Если же данных мало, и ммп равна 0.45, получаем $0.55 * (1 - 0.45) + 0.45 * 0.45 = 0.505$ и $0.55 * 1 + 0.45 * 0 = 0.55$, качество ухудшилось.

Задача 2 (1.5 балла). Метрические методы, kNN, проклятие размерности..
Вероятность, что расстояние до ближайшего соседа больше r, равна

$$1 - (1 - r^D)^N$$

Для медианы это значение равно 1/2, получаем $r = (1 - (1/2)^{1/N})^{1/D}$

При $N = 500, D = 10$ медиана примерно равна 1/2. При дальнейшем увеличении размерности пространства, получаем следующую картину:

Проклятие размерности заключается в экспоненциальной зависимости размера выборки от размерности данных для совершения статистически значимых выводов. Приведенная формула для медианы показывает, что при увеличении размерности расстояние до ближайшего соседа асимптотически приближается к максимальному, значит все объекты становятся равноудалены от центра и различия между ними теряются.

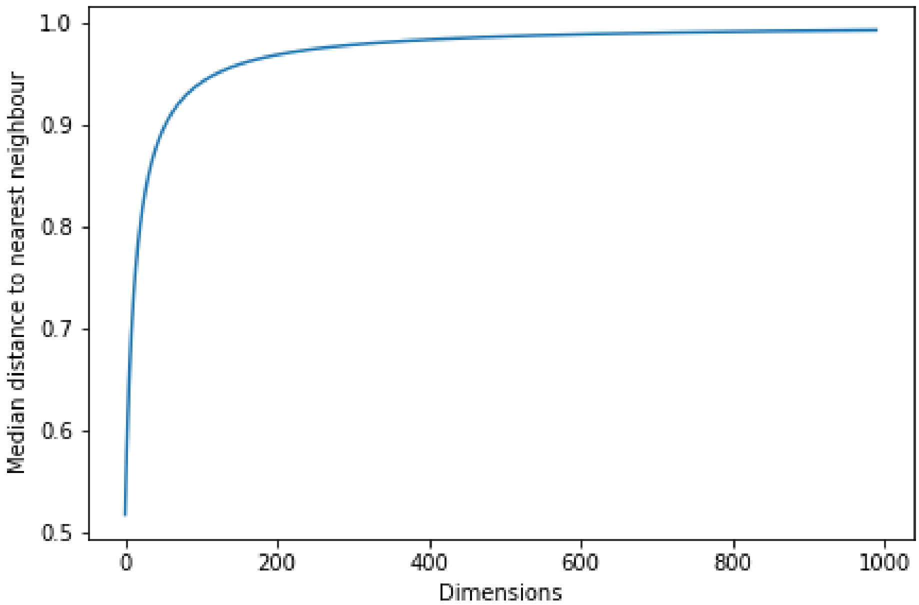

Рис. 1: Curse of dimensionality

Для того, чтобы побороть проклятие, придется собирать выборки следующих размеров в зависимости от размерности:

$$N = \frac{\ln(1/2)}{\ln(1 - (1/2)^D)}$$

Задача 3 (1 балла). Решающие деревья, индекс Джини.

1. Матожидание частоты ошибок классификатора $a(x)$ на R_m равно

$$E = \sum_k P[class(x) = k]P[a(x) \neq k] = \sum_k p_{mk}(1 - p_{mk}) = G_m$$

2. Выборочная дисперсия класса k равна

$$S_k = \frac{1}{|R_m|} \sum_{x_i \in R_m} [y_i = k]^2 - \left(\frac{1}{|R_m|} \sum_{x_i \in R_m} [y_i = k] \right)^2 =$$

$$= \frac{1}{|R_m|} \sum_{x_i \in R_m} [y_i = k] - \left(\frac{1}{|R_m|} \sum_{x_i \in R_m} [y_i = k] \right)^2 =$$

$$= p_{mk} - p_{mk}^2 = p_{mk}(1 - p_{mk})$$

Сумма дисперсий по всем классам равна

$$\sum_k p_{mk}(1 - p_{mk}) = G_m$$

Задача 4 (1.5 балла). LDA

В модели LDA, с предположением одинаковой матрицы ковариации для классов и нормального распределения классов, получаем:

$$\log\left[\frac{P(y=0|x)}{P(y=1|x)}\right] = \log\left[\frac{P(y=0)P(x|y=0)}{P(y=1)P(x|y=1)}\right] =$$

$$= \log\left[\frac{P(y=0)}{P(y=1)}\right] + \sum_{i=1}^{d} \log\left[\frac{P(x^i|y=0)}{P(x^i|y=1)}\right] =$$

$$= \log\left[\frac{P(y=0)}{P(y=1)}\right] + \sum_{i=1}^{d} \log\left[\frac{\exp\frac{-(x^i-\mu_{0,i})^2}{2\sigma_i^2}}{\exp\frac{-(x^i-\mu_{1,i})^2}{2\sigma_i^2}}\right] =$$

$$= \log\left[\frac{P(y=0)}{P(y=1)}\right] + \sum_{i=1}^{d} \frac{(x^i-\mu_{1,i})^2-(x^i-\mu_{0,i})^2}{2\sigma_i^2} =$$

$$= \log\left[\frac{P(y=0)}{P(y=1)}\right] + \sum_{i=1}^{d} \left(\frac{\mu_{0,i}-\mu_{1,i}}{\sigma_i^2}x^i + \frac{\mu_{1,i}^2-\mu_{0,i}^2}{2\sigma_i^2}\right) =$$

$$= a_0 + a^T x$$

Где a_0 и a зависят от частоты классов, матожидания и ковариации распределений $p(x|y)$ - так же, как для логистической регрессии. Но все же в общем случае эти две модели дают разные ответы, хотя бы потому что логистическая регрессия не предполагает нормального распределения классов, и в случае, когда выброс в данных приводит к одинаковой оценке средних по классам, LDA ломается, но регрессия дает адекватный ответ. Если же предположения модели LDA (многомерная нормальность, гомоскедастичность) оказываются разумными для данной задачи, то она становится предпочтительнее регрессии.

Задача 7 (1.5 балл) SVD. Для двух заданных матриц A и B одного размера найдите ортогональную матрицу Q, для которой норма Фробениуса разности $||QA-B||_F$ минимальна.

Напомним, что норма Фробениуса определяется, как

$$X_F = \sqrt{\sum_{i,j} x_{ij}^2}$$

Эту задачу можно решать по-разному, но наиболее эффективное решение использует сингулярное разложение (а какой именно матрицы — вам предстоит выяснить самим))

$$argmin_Q||QA-B||_F = argmin_Q\sqrt{tr(QA-B)^T(QA-B)} =$$

$$= argmin_Q tr(QA-B)^T(QA-B) = argmin_Q tr(A^TQ^T-B^T)(QA-B) =$$

$$= argmin_Q tr(A^T A - B^T QA - A^T Q^T B + B^T B) =$$

$$= argmax_Q tr(B^T QA + A^T Q^T B) = argmax_Q tr(QAB^T)$$

Заметим, что если матрица U ортогональна, то суммы столбцов U меньше или равны 1, а D - диагональная матрица из положительных элементов, то $trDU = \sum_i d_i \sum_j u_{ij} \leq \sum d_i$. Тогда если $AB^T = UDV$, то $tr(QAB^T) \leq trD$, и максимум достигается при $Q = V^T U^T$.