Анализ кривой обучения

Неделько В. М.

Институт математики СО РАН, г. Новосибирск
nedelko@math.nsc.ru

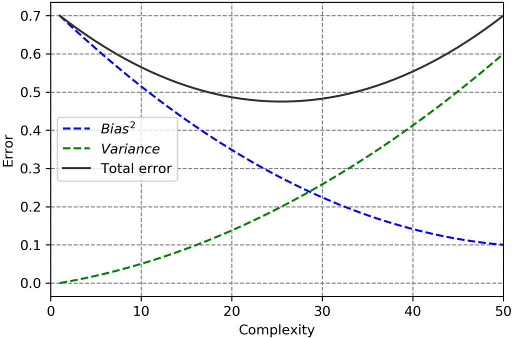
Спецкурс «Теория статистических решений».
Лекция 9.

Разложение критерия качества решающих функций

Есть (как минимум) два подхода к разложению критерия качества решающих функций.

- Разложение на смещение и разброс (bias-variance decomposition).
- Разложение на меру адекватности и меру статистической устойчивости (Г. С. Лбов, Н. Г. Старцева, 1989)

Желаемое поведение разложения



- смещение $\neq$ ошибка аппроксимации
- разброс $\neq$ ошибка статистического оценивания

Смещение и разброс

Для независимых случайных величин u и v имеет место

$$\mathsf{E}(u-v)^2 = \mathsf{D}u + (\mathsf{E}u - \mathsf{E}v)^2 + \mathsf{D}v$$

Зафиксируем точку x признакового пространства и подставим $u = y\,|\,x, \quad v = f(x)$.
Усреднение ведётся по условному распределению на целевой переменной и по выборкам заданного объёма.

Получаем «шум», смещение и разброс (в точке x).

Логарифмическая функция потерь

Логарифмическая функция потерь

$$L(y, u) = -\ln u_y$$

Критерий эквивалентен функции правдоподобия выборки.

Можно свести к дивергенции Кульбака-Лейблера.

$$L(v, u) = \sum_{\omega=1}^{k} v_\omega \ln \frac{v_\omega}{u_\omega}.$$

Универсальное разложение

Имеет место

$$\mathsf{E}L(y,u) = H(p) + L(p,\overline{u}) + \mathsf{E}L(\overline{u},u)$$

Здесь $\overline{u}$ – «главное предсказанние», но на самом деле это распределение, при котором прогноз оптимален.

Разложение для kNN

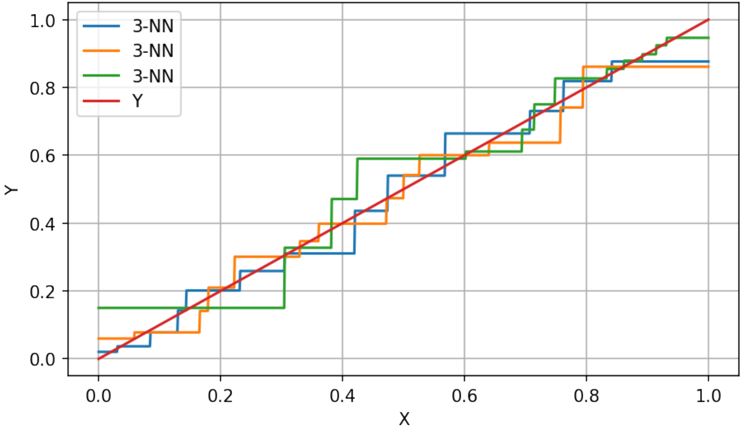
В ряде источников (напр., Hastie... The Elements of Statistical Learning) приводится следующая формула разложения для метода kNN

$$\mathsf{E}(y - f(x))^2 \,|\, x = \left(f(x) - \frac{1}{k} \sum_{i=1}^{k} f(N_i(x)) \right)^2 + \frac{\sigma^2}{k} + \sigma^2.$$

В классической постановке задачи регрессии, т.е. когда координаты X в обучающей выборке фиксированы (не случайны), первое слагаемое есть bias.

В общей постановке, когда X случайны, первое слагаемое содержит как смещение, так и разброс.

kNN для одномерной регрессии



Смещение существенно только на краях области.

Разложение для kNN

Без учёта краевых эффектов имеем разложение

$$\mathsf{E}(y(x) - f(x))^2 \approx \frac{(k+1)(k+2)}{12N^2k} + \frac{\sigma^2}{k} + \sigma^2. \qquad (1)$$

Первое и второе слагаемые — оба variance.
Как видим, с ростом сложности variance сначала уменьшается и только потом растёт.

При большей размерности пространства эффект исчезает.

Меры адекватности и устойчивости

Идея подхода (Г. С. Лбов, Н. Г. Старцева, 1989) в том, чтобы разложить ошибку на погрешность аппроксимации и статистическую погрешность.
Асимптотическое значение среднего качества (риска)

$$R = \lim_{N \to \infty} \mathsf{E} L(y, u)$$

Мера адекватности есть разность между асимптотическим средним риском и байесовским уровнем ошибки.
Мера статистической устойчивости есть разность между средним риском и асимптотическим.

Отступ как мера сложности

В качестве универсальной меры сложности решения можно использовать отступ (ненормированный).

Нормированный отступ фигурирует в оценках обобщающей способности.
Неномированный отступ характеризует степень обучения (и переобучения).

Отступ вычисляется на основе оценок вероятности.

Процесс обучения

- Ошибка на обучении монотонно уменьшается (если это не так, значит обучение «пошло вспять»).
- Иногда можно обучаться и после достижения нулевой ошибки.
- По кроссвалидации можно оптимизировать предельное значение сложности, а в рамках порога выбирать минимум на обучении.

Выводы

Результаты исследования (проведённые численно и аналитически) показывают, что в ряде примеров как смещение, так и разброс, могут вести себя немонотонно при увеличении сложности (сначала уменьшаться, а затем расти).

В следствие этого разложение на смещение и разброс не является удовлетворительным объяснением процесса обучения, поскольку смещение и разброс лишь косвенно характеризуют аппроксимационную способность и статистическую устойчивость метода построения решающих функций.

В качестве универсальной меры сложности решения можно использовать отступ (ненормированный).