

Аналитические выражения для разложения ошибки метода kNN на смещение и разброс

Неделько Виктор Михайлович

Институт математики им. С.Л. Соболева

Международная конференция ИОИ-14. Россия, г. Москва,
6–9 декабря 2022 г.

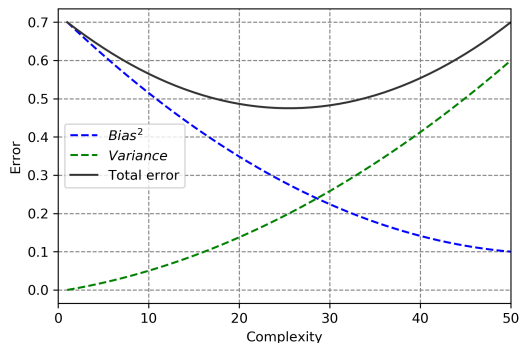
Goals

When choosing the optimal complexity of the method for constructing decision functions, an important tool is the decomposition of the quality criterion into bias and variance.

Here we obtain an expression for the variance component for the kNN method for the linear regression problem in the formulation when the “explanatory” features are random variables.

In contrast to the well-known result obtained for non-random “explanatory” variables, in the considered case, the variance may increase with the growth of k .

Desired properties of decomposition



This is not always the case.

- Bias \neq approximation error
- Variance \neq stochastic error

Regression problem

Let X be the space of values of variables used for forecasting, and Y be the set of values of the predicted variable.

All variables are random variables with some joint distribution function.

Decision function is a mapping $f : X \rightarrow Y$.

The decision function is constructed based on some training sample of size N

$$S_N = ((x^\omega, y^\omega), \omega = \overline{1, N}).$$

For the decision function as a whole, the quality criterion will be MSE, i.e.

$$R(f(\cdot)) = \mathbf{E}_{x,y}(y - f(x))^2.$$

By this criterion, the optimal solution will be a regression function, i.e. a conditional mathematical expectation.

Classical statement

In the classical statement of regression problem, the values of X are not random. Only the target variable is random, which is represented as

$$y(x) = \hat{f}(x) + \delta, \quad (1)$$

where $\hat{f}(x)$ is some unknown function, and δ is a random variable with zero mean and variance σ^2 .

Bias-Variance decomposition

For arbitrary independent random variables u and v (if the corresponding moments exist), the identity holds

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

where \mathbf{D} denotes variance, i.e. $\mathbf{D}u \equiv \mathbf{E}u^2 - (\mathbf{E}u)^2$.

Bias-Variance decomposition

Let's fix a point x of the feature space and substitute $u = y | x$, $v = f(x)$. Since $f(x)$ is constructed on a random sample, v is a random variable. Then we get

$$\begin{aligned} \mathbb{E}_{S_N, y|x} (y - f(x))^2 = \\ \mathbb{D}_{y|x} y + (\mathbb{E}_{y|x} y - \mathbb{E}_{S_N} f(x))^2 + \mathbb{D}_{S_N} f(x). \end{aligned} \quad (2)$$

The notation $\mathbb{E}_{S_N, y|x}$ means that the expectation is taken over all samples of size N and over the conditional distribution on the target variable y at the point x . So, a subscript at operators \mathbb{E} or \mathbb{D} indicates the domain for averaging.

We obtain that in this formulation is the decomposition of MSE into “noise”, bias and variance.

Decomposition for kNN

A number of sources (e.g. Hastie... The Elements of Statistical Learning) provide the following decomposition formula for the kNN method

$$\mathbb{E}_{S_{N,y}|x} (y - f(x))^2 = \left(f(x) - \frac{1}{k} \sum_{i=1}^k \hat{f}(\xi_i(x)) \right)^2 + \frac{\sigma^2}{k} + \sigma^2, \quad (3)$$

where $\xi_i(x)$ is the coordinates of the i -th “neighbor” of a point x . The second term in this decomposition is proposed to be interpreted as a variance.

The variance component in 3 decreases monotonically with the growth of k , i.e. it increases with increasing complexity, since the complexity characteristic for kNN is opposite to k and can be, for example, $\frac{1}{k}$.

Random features case

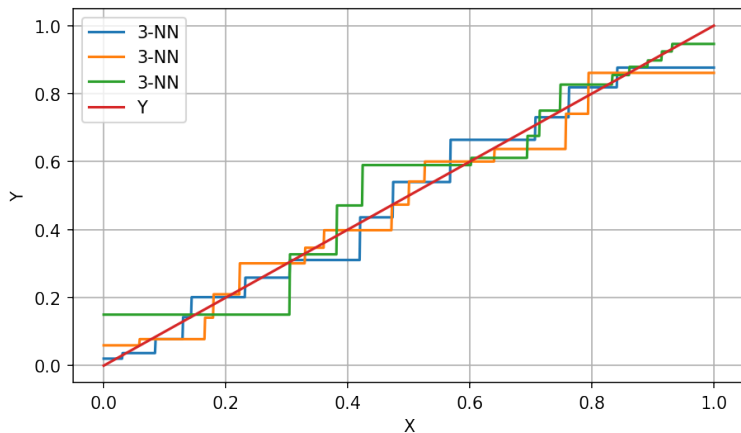
Consider now “explanatory” features to be random.

Let $X = [0, 1]^n$ and $y = x_1 + \delta$, where $x = (x_1, \dots, x_n) \in X$.

We consider the model $\hat{f}(x) = x_1$ as a linear regression model without loss of generality because any linear model may be converted to it via proper transformation of features.

Suppose that x_j are independent random variables, $x_j \sim U(0, 1)$.

kNN regressions for different samples



The bias is close to 0.

Decomposition for the random features case

Proposition. For the inner points of X , there is:

$$\mathbb{E}_{S_N, y|x} (y - f(x))^2 = \mathbb{D} \left[\frac{1}{k} \sum_{i=1}^k \xi_i(x) \right] + \frac{\sigma^2}{k} + \sigma^2, \quad (4)$$

where

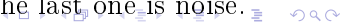
$$\mathbb{D} \left[\frac{1}{k} \sum_{i=1}^k \xi_i(x) \right] \approx \frac{(N^*)^{-\frac{2}{n}}}{2n^2 k^2} \sum_{m=0}^{k-1} \frac{k-m}{m!} \Gamma \left(m + \frac{2}{n} \right) \quad (5)$$

and

$$N^* = NV_0, \quad V_0 = \frac{\pi^{\frac{n}{2}}}{2^n \Gamma \left(1 + \frac{n}{2} \right)}.$$

The decomposition is asymptotically exact as $N \rightarrow \infty$.

Here V_0 is the volume of n -dimensional ball of diameter 1.

The first two terms in 4 are the variance, the last one is noise. 

Special case

For $n = 1$ the formula 5 get

$$D \left[\frac{1}{k} \sum_{i=1}^k \xi_i(x) \right] = \frac{(k+1)(k+2)}{12N^2k}.$$

In contrast to 1, the resulting decomposition has a monotonically increasing (close to linear growth) term in the variance component.

This term provides the possibility of decreasing variance with increasing complexity.

Properties

For $n = 2$ the formula 5 get simple

$$D \left[\frac{1}{k} \sum_{i=1}^k \xi_i(x) \right] = \frac{k+1}{4\pi Nk}.$$

We can see that the variance for kNN demonstrates different behavior depending on dimensionality.

- By $n = 1$ the variance increases as k increases.
- By $n = 2$ the variance tends to a positive constant as $k \rightarrow \infty$.
- By $n > 2$ the variance tends to zero.

«Обоснование» эффективности случайного леса

Сравниваем по точности три варианта решающей функции.

- Одно «обычное» дерево.
- Одно «плохое» дерево (построено по искажённой выборке).
- Ансамбль «плохих» деревьев (случайный лес).

Легко доказать, что 3 лучше чем 2, но ниоткуда не следует, что 3 лучше 1, поскольку смещение и разброс в 2 больше чем в 1.

«Обоснование» эффективности бустинга

Сравниваем по точности три варианта решающей функции.

- Одно «обычное» дерево.
- Одно дерево, которое вообще не годится в качестве самостоятельного решения (не прогнозирует целевую переменную, а «исправляет ошибки» предыдущих).
- Ансамбль «несамостоятельных» деревьев (boosting).

Легко поверить, что 3 лучше чем 2, но ниоткуда не следует, что 3 лучше 1, поскольку смещение в 2 больше чем в 1 и деревья зависимы.

«Обоснование» эффективности композиций

Текущее состояние вопроса.

- Композиции как правило действительно эффективнее отдельных методов.
- Разложение на смещение и разброс вероятно отражает некоторую интуицию, стоящую за этим фактом.
- Однако, данная конструкция не позволяет *доказать*, что композиции и должны быть эффективнее.

Alternative decomposition

Such undesired properties of bias-variance decomposition encourage to search alternatives.

As such alternative might be considered another decomposition of the error: into a measure of adequacy and a measure of stability, that was proposed by G.S. Lbov and N.G. Startseva (Complexity of Distributions in the Classification Problem. *Doklady RAS*, 1994, vol. 338, no. 5, pp. 592–594).

The idea of the approach is to decompose the error into the approximation error and the statistical error.

The components of this decomposition are obviously monotonic.

Adequacy and stability measures

The idea behind the approach is to decompose the error into an approximation error and a statistical error.

The more complex class of decision functions, the more accurately it can potentially approximate optimal solution, but the actual accuracy may decrease due to statistical error when solution is built on sample.

The basic concept of this decomposition is the asymptotic average risk or the asymptotic value of the average quality

$$F_{\infty}(Q) = \lim_{N \rightarrow \infty} F_N(Q). \quad (6)$$

Adequacy and stability measures

The measure of adequacy is the difference between the asymptotic mean risk and the Bayesian risk. This measure shows how good a solution the method could give in the case of an unlimited sample (or when constructing solutions on the distributions themselves).

The measure of statistical stability is the difference between the average risk and the asymptotic one.

Approximational and stochastic errors

Note that it would be more correct to call the introduced components a measure of inadequacy (approximational error) and a measure of instability (statistical error), since they characterize error rather than accuracy. To avoid terminological inconveniences, we will also use the terms approximation error and statistical estimation error.

The Bayesian level of error (risk) is exactly what in the equation 2 called noise.

We see that the decompositions are similar: both have three components, one of which (noise) coincides.

Learning on distributions

To estimate the measure of adequacy we need to learn model on distributions instead of samples.

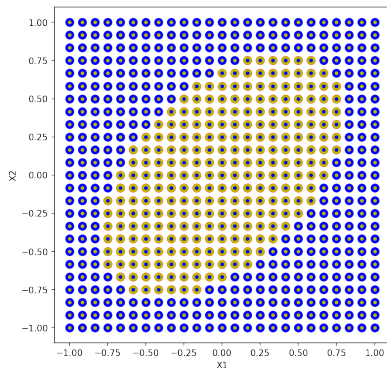
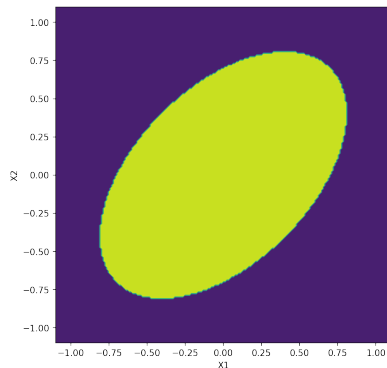
All methods (that I know) can use distributions instead of sample.

For SVM this is not obvious, but SVM is equivalent to some unconditional optimization.

Now we show some examples for the gradient boosting been learned on distribution, when the probabilities were expressed by sample weights.

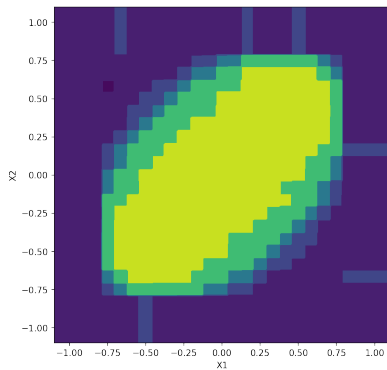
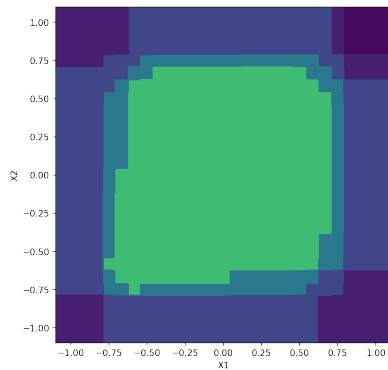
Usually, gradient boosting can overtrain infinitely, but in the cases below it converges.

Regular weighted “sample” as a distribution



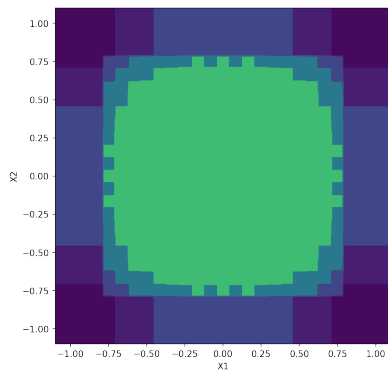
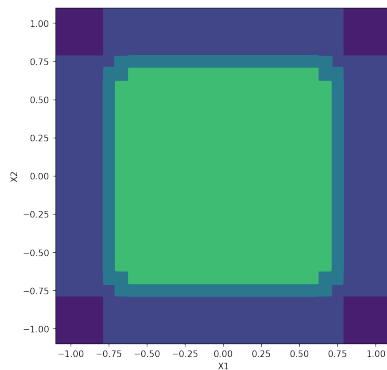
$P(y = 1 | x \in E) = P(y = 0 | x \notin E) = 0.9$, E is the ellipse,
 $x_j \sim U(-1, 1)$.

Gradient boosting solution



$\text{max_depth} = 2$, $n_estimators$ is 15 and 1000 correspondently.

Gradient boosting on stumps



`max_depth = 1`, `n_estimators` is 15 and 1000 correspondently.

Reference

Nedel'ko V. M. On Decompositions of Decision Function Quality Measure // The Bulletin of Irkutsk State University. Series Mathematics, 2020, vol. 33. — p. 64–79.
<https://doi.org/10.26516/1997-7670.2020.33.64>

Conclusions

- We obtain an expression for the variance component for the kNN method for the linear regression problem in the formulation when the “explanatory” features are random variables.
- Comparison of two decompositions was carried out: for bias and variance and for measure of adequacy and stability.
- It is shown that with increasing complexity, the bias can increase, and the variance can decrease, while the decomposition into the measure of adequacy and stability always has a “canonical” form.
- Non-trivial observations can be made when training models on distributions.