Predicting Market Openings using LOOCV Cross-Validation k-Nearest Neighbors with Asymptotically Optimal Weights

Alejandro Moral Aranda

January 31, 2025

Abstract

We propose an application of cross-validated weighted k-nearest neighbors (k-NN) with asymptotically optimal weights to forecast whether a market will open in positive or negative territory. Our method leverages closing prices and other signals from adjacent markets—such as major currency pairs, Bitcoin, and potentially equity index futures—to capture global intermarket relationships that can influence U.S. equity market openings. The asymptotically optimal weights in the k-NN classifier reduce misclassification risk, while cross-validation ensures robust parameter selection, mitigating overfitting. Experimental results will assess the model's accuracy, stability, and practical utility in short-term market forecasting.

LOOCV for Weighted k-NN with Asymptotically Optimal Weights

1. Dataset. Let

$$\mathcal{D} \;=\; \{(X_i, Y_i) \mid X_i \in \mathbb{R}^d,\; Y_i \in \{1,2\},\; i = 1, \ldots, n\}$$

be an i.i.d. sample from an unknown distribution on $\mathbb{R}^d \times \{1,2\}$.

2. Weighted k-NN Classifier. For a test point $x \in \mathbb{R}^d$, reorder the sample points (excluding a point if we are doing leave-one-out) by increasing distance to x:

$$\|X_{(1)} - x\| \;\le\; \|X_{(2)} - x\| \;\le\; \cdots \;\le\; \|X_{(n)} - x\|.$$

Given a chosen integer k and weights $\{w_{ni}\}_{i=1}^{n}$, the *weighted* k-nearest-neighbor decision rule assigns:

$$C(x) \;=\; \begin{cases} 1, & \text{if } \displaystyle\sum_{i=1}^{k} w_{ni}\, \mathbf{1}\{\, Y_{(i)} = 1\} \;\ge\; \tfrac{1}{2}, \\ 2, & \text{otherwise.} \end{cases}$$

Here, $Y_{(i)}$ denotes the label of the i-th closest neighbor. Let n be the sample size and d the feature dimension. Define

$$k^* \;=\; \Big\lfloor B^*\, n^{\frac{4}{d+4}} \Big\rfloor,$$

For $1 \le i \le k^*$, the *asymptotically optimal weight* w_{ni}^* assigned to the ith nearest neighbor (ranked by distance) is

$$w_{ni}^* \;=\; \frac{1}{k^*}\Big[1 \;+\; \frac{d}{2} \;-\; \frac{d}{2\,(k^*)^{2/d}}\Big]\Big(i^{1+\frac{2}{d}} \;-\; (i-1)^{1+\frac{2}{d}}\Big)^{-1},$$

and for $i = k^* + 1, \ldots, n$, we set

$$w_{ni}^* \;=\; 0.$$

The weights may be normalized so that

$$\sum_{i=1}^{n} w_{ni}^* \;=\; 1.$$

3. Leave-One-Out Cross-Validation (LOOCV). For each $j = 1, \ldots, n$, define the *LOOCV classifier* C_{-j} that excludes the j-th observation (X_j, Y_j)

from the training set. That is, we find the neighbors of X_j only among $\{(X_i, Y_i) : i \neq j\}$. Then,

$$C_{-j}(X_j) \;=\; \begin{cases} 1, & \text{if } \sum_{i=1}^{k} w_{ni}\, \mathbf{1}\{\, Y_{(i)}^{(-j)} = 1\} \;\geq\; \frac{1}{2}, \\ 2, & \text{otherwise,} \end{cases}$$

where $\{(X_{(1)}^{(-j)}, Y_{(1)}^{(-j)}), \ldots, (X_{(n-1)}^{(-j)}, Y_{(n-1)}^{(-j)})\}$ is the training set $\mathcal{D} \setminus \{(X_j, Y_j)\}$ reordered by distance to X_j.

4. Error and Hyperparameter Selection. The *LOOCV error* for a fixed choice of $(k, \{w_{ni}\})$ is

$$\mathrm{LOOCV}(k, \{w_{ni}\}) \;=\; \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}\Big\{C_{-j}(X_j) \;\neq\; Y_j\Big\}.$$

Finally, we select

$$\big(k^*, \{w_{ni}^*\}\big) \;=\; \underset{k,\, \{w_{ni}\}}{\arg\min}\, \mathrm{LOOCV}\big(k, \{w_{ni}\}\big).$$

In other words, we choose both the number of neighbors k and the weight sequence $\{w_{ni}\}$ that minimize the average leave-one-out misclassification rate.

5. Using the k Neighbors for Numerical Forecasts. After deciding the class label (e.g., whether the market opens positive or negative), one may further exploit the same k nearest neighbors for a *quantitative* prediction. Suppose each point (X_i, Z_i) in the dataset has an associated real-valued variable Z_i (e.g., size of the opening gap). Once k neighbors of x are identified:

- **Weighted Linear Regression:** Fit a linear model on the k neighbors via weighted least squares.

$$\min_{\beta_0, \beta_1, \ldots, \beta_d} \sum_{i=1}^{k} w_{ni}^* \Big(Z_{(i)} - \beta_0 - \beta_1\big(X_{(i),1} - x_1\big) - \cdots - \beta_d\big(X_{(i),d} - x_d\big)\Big)^2,$$

  yielding a local linear approximation around x.

By coupling the nearest-neighbor classification decision with a local regression approach, one can not only infer the likely direction (positive vs. negative) but also estimate the magnitude of the market move, offering a richer predictive framework for short-term financial forecasting.

Asymptotically Optimal Weight Formula (Samworth, 2012)

1 Sample and Summary Statistics

Asymptotically Optimal Weights, Samworth2012

A set of weights for a k-nearest neighbor (k-NN) classifier is called *asymptotically optimal* if, under suitable smoothness and regularity conditions on the data-generating process (particularly on the regression function $\eta(x) = \mathbb{P}(Y = 1 \mid X = x)$), these weights minimize the leading term in the misclassification probability as the sample size $n \to \infty$.

Concretely, let $X \in \mathbb{R}^d$ have a distribution with density measures P_1 and P_2 conditional on $Y = 1$ and $Y = 2$, respectively. Let $\pi = \mathbb{P}(Y = 1)$. Then, for each test point $x \in \mathbb{R}^d$, we order the training samples $(X_{(1)}, Y_{(1)}), \ldots, (X_{(n)}, Y_{(n)})$ by distance to x. We choose

$$k^* \;\approx\; B^* \, n^{\frac{4}{d+4}}$$

for some distribution-dependent constant B^*, and assign weights w^*_{ni} to these k^* neighbors according to a formula that prioritizes closer neighbors while tapering off smoothly for those ranked further away.

These weights ensure the classifier achieves a minimax-optimal rate of convergence for the excess risk (i.e., the probability of misclassification above the Bayes error) and thereby is *asymptotically optimal* among all weighted k-NN schemes.

References

[1] R. J. Samworth (2012). *Optimal Weighted Nearest Neighbour Classifiers.* The Annals of Statistics, 40(5), 2733–2762.