Stat 205: Introduction to Nonparametric Statistics

Lecture 10: Nearest Neighbors (Theory)

Instructor David Donoho; TA: Yu Wang

k-nn Setting

- Data $Y = (y_i)_{i=1}^n$; "response", "target"
  - Regression: y_i continuous response variable
  - Classification: y_i categorical [eg binary] response variable,$\{1, \dots, C\}$
- Data $X = (x_i)_{i=1}^n$; $x_i \in \mathbf{R}^p$; predictors.
- Examples
  - Credit card fraud:
    - $y_i \in \{0, 1\}$ 1=legit/0=fraud
    - $x_i = (x_{i,1}, x_{i,2})$ eg

$$\begin{aligned} x_{i,1} &= \#\{\text{previous dollars spent at similar merchant}\} \\ x_{i,2} &= \#\{\text{previous dollar purchases of similar item }\} \end{aligned}$$

  - Reservoir Permeability [Example 4.6 in Wasserman]
    - y_i rock permeability
    - $x_{i,1}$ area of pore spaces
    - $x_{i,2}$ perimeter of pore spaces

k-Nearest-neighbor Theory

- **Generative model**

$$y_i \sim p(y|x_i), \qquad i = 1, \ldots, n.$$

  - Regression: $y = \mu(x) + z$;
    $E(z|x) = 0$; for example $z \sim N(0, 1)$.
  - Classification: $p(y|x)$ a discrete probability on $\{1, \ldots, C\}$.
- **Performance:** aka "Risk". Two standard options
  - Regression: $PMSE(m, x) = \mathbb{E}\left[(y - m(x))^2 | y \sim p(y|x)\right]$.
  - Classification: $PErr(m, x) = \Pr\left(y \neq m(x) | y \sim p(y|x)\right)$.
- **Optimality:**
  - Regression: $\mu(x|Y) = \mathbb{E}\left[y|x\right]$.
  - Classification: $\gamma(x|Y) = \text{argmax}_c \Pr\left(y = c|x\right)$.

k-Nearest-neighbor Procedure

- $d(x, x')$ 'distance' between p-dimensional feature vectors
- Tuning Parameter: k number of neighbors
- For given x, $N_k(x)$ is the set of k-nearest neighbors

$$N_k(x) = \{i : d(x_i, x) \text{ is among the } k \text{ smallest distances } d(x_j, x)\}$$

  (assume no ties, or if ties break randomly)
- k-nn Estimator [regression]
$$\hat{\mu}^{knn}(x) = Ave\{y_i | x_i \in N_k(x)\}.$$
- k-nn Estimator [classification]
$$\hat{\gamma}^{knn}(x) = \text{argmax}_c \{y_i = c | x_i \in N_k(x)\}.$$
- Heuristic: find some nearby examples, summarize them, decide
- Humans use this principle. (Robert Cialdini, Influence)

Theory in noiseless special case, 1

Simplest case (**no noise**):

- $k = 1$ nearest neighbor
- $N(x) = N(x; X)$ index of 1-nearest neighbor of x within X
- $x', x_1, \ldots, x_n \sim_{iid} F$.
- Regression $y_i = \mu(x_i)$
  Error:
  $$\mu(x) - y_{N(x)} = \mu(x) - \mu(x_{N(x)})$$
  Risk
  $$PMSE = \mathbb{E}\left[(\mu(x') - \mu(x_{N(x')}))^2\right]$$
- Classification $y_i = \gamma(x_i)$
  Error:
  $$\{\gamma(x') \neq y_{N(x')}\} = \{\gamma(x) \neq \gamma(x_{N(x)})\}$$
  Risk
  $$PErr = \Pr\left(\gamma(x') \neq \gamma(x_{N(x')})\right)$$

Theory in noiseless special case, 2

When does $Risk \to 0$ as $n \to \infty$?

- **Needed Fact (next slide)** Nearest Neighbor distance: $X^{(n)}$ dataset with n observations, F has no discrete components.

$$d(x', x_{N(x';X_n)}) \to 0, \qquad n \to \infty; \qquad x \in Support(F).$$

- $\mathcal{X}_\mu = \{$ continuity points $\}$ of μ [resp .,γ]

$$\mu(x') \to \mu(x), \quad \text{resp } \gamma(x') \to \gamma(x) \qquad d(x', x) \to 0.$$

- Consequences of Continuity:
  - Regression:
  $$PMSE(\hat{\mu}_n^{1nn}(x)|x) \to 0, \qquad x \in \mathcal{X}_\mu.$$
  - Classification:
  $$PErr(\hat{\gamma}_n^{1nn}(x)|x) \to 0, \qquad x \in \mathcal{X}_\mu.$$
- Suppose that for $x' \sim F$, $P(x' \in \mathcal{X}_\mu) = 1$ (piecewise continuity in probability) $\mathbb{E}\left[\mu^2(x')\right] < \infty$ (finite prediction variance)

$$\begin{aligned} PMSE_n &= \mathbb{E}\left[PMSE(\hat{\mu}_n^{1nn}(x')|x')\right] \to 0. \\ PErr_n &= \mathbb{E}\left[PErr(\hat{\gamma}_n^{1nn}(x')|x')\right] \to 0. \end{aligned}$$

- Heuristics:
  - knn regression works well if $\mu(\cdot)$ piecewise continuous.
  - knn classifier works well if $\gamma(\cdot)$ piecewise continuous.

Theory in noiseless special case, 3

Needed Fact: When does nearest neighbor converge?

- Nearest Neighbor distance: $X^{(n)}$ dataset with:
  - n observations, $x_i \sim_{iid} F, \qquad i = 1, \dots, n$
  - Fix $\delta > 0$; let $x \sim F$ and define

$$p(\delta|x') \equiv P(d(x', x) \leq \delta)$$

- **Theorem.** *Suppose* $p(\delta|x') > 0$ *whenever* $\delta > 0$*; then*

$$d(x', x_{N(x';X_n)}) \to 0, \qquad n \to \infty;$$

- **Proof.**
  - Fix $\delta > 0$

$$\{d(x', x_{N(x';X_n)}) > \delta\} = \cap_{i=1}^{n} \{d(x', x_i) > \delta\}$$

  - Independence allows product rule:

$$P\{d(x', x_{N(x';X_n)}) > \delta\} = [P\{d(x', x_i) > \delta\}]^n$$

  - Denote $\varepsilon = P(d(x', x) \leq \delta) > 0$

$$[P\{d(x', x_i) > \delta\}]^n = [1 - \epsilon]^n = \exp(n \log(1 - \varepsilon)) \to 0.$$

- **Meaning.**
  - NN converges within *support* of F
  - Don't expect NN to converge *outside support* i.e. don't extrapolate.
  - Support $X_F \equiv \{x' : p(\delta|x') > 0\}$
  - Example: $x \sim F \equiv N(\mu, \Sigma)$ on $\mathbf{R}^p$;

$$p(\delta|x') \sim f(x')\delta^p \text{ as} \delta \to 0.$$

  Support $= X_F = \mathbf{R}^p =$ 'everything'.

Theory in noiseless special case, 4

Nearest Neighbor Methods
Generalities
1nn, no noise
1nn under Noise
Regression
Classification
knn

1NN Noiseless Regression Theorem. *Suppose RV* $x' \sim F$,

- *Continuity points of* $\mu(x)$:
$$\mathcal{X}_\mu = \{x : \mu(x) = \lim_{x' \to x} \mu(x')\}$$
- *Piecewise continuity in probability:*
$$P(x' \in \mathcal{X}_\mu) = 1$$
- *Finite variance of predictor:*
$$\mathbb{E}\left[\mu^2(x')\right] < \infty.$$

$$PMSE_n \quad = \quad \mathbb{E}\left[PMSE(\hat{\mu}_n^{1nn}(x')|x')\right] \to 0.$$

knn regression works well if $\mu(\cdot)$ piecewise continuous.

Theory in noiseless special case, 5

Continuity points of $\gamma(x)$:

$$\mathcal{X}_\gamma = \{x : \gamma(x) = \lim_{x' \to x} \gamma(x'))\}$$

1NN Noiseless Classification Theorem.

Suppose

- RV $x' \sim F$,
- Piecewise continuity in probability:

$$P(x' \in \mathcal{X}_\gamma) = 1$$

- Then

$$\begin{aligned} PErr_n &= \mathbb{E}\left[PErr(\hat{\gamma}_n^{1nn}(x')|x')\right] \\ &= \Pr\left(y_{N(x')} \neq y_{x'}\right) \\ &= \Pr\left(\gamma(x_{N(x')}) \neq \gamma(x')\right) \\ &\to 0. \end{aligned}$$

knn classifier works well if $c(\cdot)$ piecewise continuous.

1nn regression noisy case, 1

- Continuity points of μ:

$$\mathcal{X}_\mu = \{x : \mu(x) = \lim_{x' \to x} \mu(x')\}$$

1NN Noisy Regression Theorem. Suppose

- RV $x' \sim F$,
- Piecewise continuity in probability:

$$P(x' \in \mathcal{X}_\mu) = 1$$

Then, as $n \to \infty$,

$$\begin{aligned} PMSE_n^{1nn} &= \mathbb{E}\left[PMSE(\hat{\mu}_n^{1nn}(x')|x')\right] \\ &= \mathbb{E}\left[(y' \neq y_{N(x')})^2\right] \\ &\to 2 \cdot OMSE. \end{aligned}$$

here OMSE denotes the best achievable MSE, using $\mu(x')$.

knn regression within factor 2 of optimal if $\mu(\cdot)$ piecewise continuous

1nn regression noisy case, 2

- Regression $y_i = \mu(x_i) + z_i$; $x', x_1, \ldots, x_n \sim_{iid} F$ $z', z_i \sim_{iid} N(0, \sigma^2)$ (say)
- Error:

$$y' - \hat{\mu}^{1nn}(x') = [\mu(x') + z'] - [\mu(x_{N(x')}) - z_{N(x')}]$$

  Risk

$$\begin{aligned} PMSE^{1nn}_{noise} &= \mathbb{E}\left[([\mu(x') + z' - [\mu(x_{N(x')}] + z_{N(x')}])^2\right] \\ &= \mathbb{E}\left[([\mu(x') - \mu(x_{N(x')})] - [z' + z_{N(x')}])^2\right] \end{aligned}$$

- $z', (z_i)$ are independent of $\{x', (x_i)\}$, so for any function $b(x', (x_i))$,

$$\mathbb{E}\left[(b(x', (x_i)) - [z' + z_{N(x')}])^2\right] = \mathbb{E}\left[b(x', (x_i))^2\right] + \mathbb{E}\left[[z' + z_{N(x')}]^2\right]$$

- By independence of $z', (z_i)$ from each other and from $\{x', (x_i)\}$,

$$\mathbb{E}\left[[z' + z_{N(x')}]^2\right] = 2\sigma^2.$$

- Set $b(x', x) = \mu(x') - \mu(x_{N(x')})$.

$$\begin{aligned} PMSE^{1nn}_{noise,n} &= \mathbb{E}\left[b^2\right] + 2\sigma^2 \\ &= \mathbb{E}\left[(\mu(x') - \mu(x_{N(x')}))^2\right] + 2\sigma^2 \\ &= PMSE^{1nn}_{nonoise,n} + 2\sigma^2. \end{aligned}$$

- Hence, by **Noiseless 1nn Theorem** (above)

$$PMSE^{1nn}_{noise,n}(x') \to 2\sigma^2, \qquad n \to \infty.$$

- Optimal Risk: $OMSE_{noise} = \mathbb{E}\left[(\mu(x') - y'))^2\right] = \sigma^2$

$$PMSE^{1nn}_{noise,n}(x') \to 2 \cdot OMSE_{noise}, \qquad n \to \infty.$$

1nn classification Noisy Case, 1

- Conditional PMF
$$q_x(c) = \Pr\left(y = c | x\right), \quad \forall c$$
- Continuity of conditional PMF:
$$q_x(c) = \lim_{x' \to x} q_{x'}(c), \quad \forall c$$
- Continuity points of q_x:
$$X_q = \{x : q_x = \lim_{x' \to x} q_{x'})\}$$

1NN Noisy Classification Theorem. Suppose

- RV $x' \sim F$,
- Piecewise continuity in probability:
$$P(x' \in X_q) = 1$$

Then, for each $\varepsilon > 0$, as $n \to \infty$,

$$\begin{aligned} PErr_n &= \mathbb{E}\left[PErr(\hat{\gamma}_n^{1nn}(x')|x')\right] \\ &= \Pr\left(y' \neq y_{N(x')}\right) \\ &\leq 2 \cdot OErr + \varepsilon. \end{aligned}$$

knn classifier works well if $c(\cdot)$ piecewise continuous.

1nn classification noisy case, 2

- Classification $y_i \sim P(y|x_i)$; $x', x_1, \ldots, x_n \sim_{iid} F$

$$\begin{aligned} PErr^{1nn}_{noise} &= \Pr\left(y' \neq \hat{\gamma}^{1nn}(x')\right) \\ &= 1 - \Pr\left(y' = \hat{\gamma}^{1nn}(x')\right) \\ &= 1 - \Pr\left(y' = y_{N(x')}\right). \end{aligned}$$

- Random PMFs $y'|x' \sim q'$; $y_{N(x';X_n)}|x_{N(x';X_n)} \sim q_n$.
- By independence of y', x' from each other and from $\{(y_i), (x_i)\}$,

$$\Pr\left(y' = y_{N(x')}\right) = \sum_c q'(c) q_n(c).$$

- **Continuity** : $\mathbb{E}\left[\|q_n - q\|_1\right] \to 0$ as $n \to \infty$.

$$\begin{aligned} PErr^{1nn}_{noise,n} &= 1 - \sum_c q'(c) q_n(c) \\ &\to 1 - \sum_c (q'(c))^2 \\ &= \Pr\left(y' \neq y''\right). \end{aligned}$$

  where y'' is iid $y'|x'$.
- Optimal Risk:

$$OErr_{noise} = \min_c \Pr\left(y' \neq c\right).$$

  Needed Fact (below)

$$\Pr\left(y' \neq y''\right) \leq 2 \min_c \Pr\left(y' \neq c\right).$$

$$PErr^{1nn}_{noise,n}(x') \leq 2 \cdot OErr_{noise} + \varepsilon, \qquad n \to \infty \quad \forall\, \varepsilon > 0.$$

Needed Fact

Lemma. *Let y', y'' be iid q'.*

$$\Pr\left(y' \neq y''\right) \leq 2 \min_c \Pr\left(y' \neq c\right).$$

Proof. Indeed, let $y' \sim q'$ and set $\gamma = \text{argmax}_c q'(c)$ (suppose unique); also

$$\alpha \equiv q'(\gamma) = max_c q'(c).$$

so

$$1 - \alpha = \min_c \Pr\left(y' \neq c\right) = \varepsilon, \text{ (say)}.$$

Now since $\sum_c q'(c)^2 \geq q'(\gamma)^2 \equiv \alpha^2$,

$$\Pr\left(y' \neq y''\right) = 1 - \sum_c q'(c)^2 \leq 1 - \alpha^2$$

and

$$1 - \alpha^2 = (1 - (1 - \varepsilon)^2) = 2\varepsilon - \varepsilon^2 \leq 2\varepsilon = 2 \cdot \min_c \Pr\left(y' \neq c\right).$$

k-NN

What happens for larger k?

- Regression: under similar asssumptions, as k increases

$$PMSE_n^{knn} \to OMSE + \delta_k$$

  where $\delta_k \to 0$ as $k \to \infty$.

- Classification: under similar asssumptions, as k increases

$$PErr_n^{knn} \to OErr + \delta_k$$

  where $\delta_k \to 0$ as $k \to \infty$.

Generally speaking there is an optimal $k_n(\{(x_i, y_i)\})$
We can use LOOCV