Statistical Machine Learning

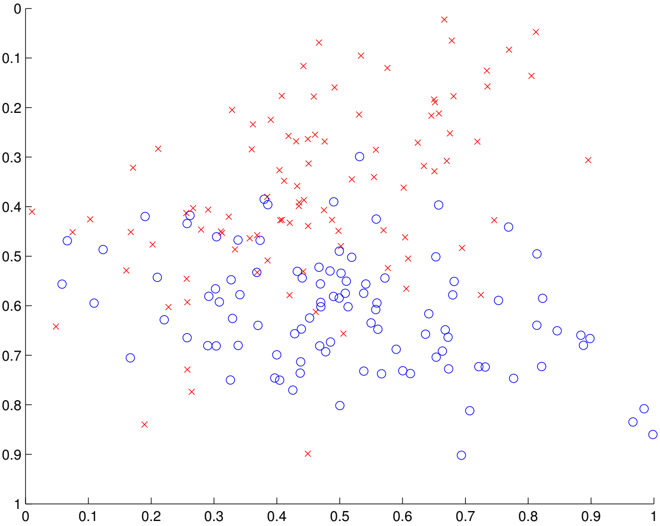
UoC Stats 37700, Winter quarter

Lecture 3: nearest neighbors and local averaging rules.

Principle

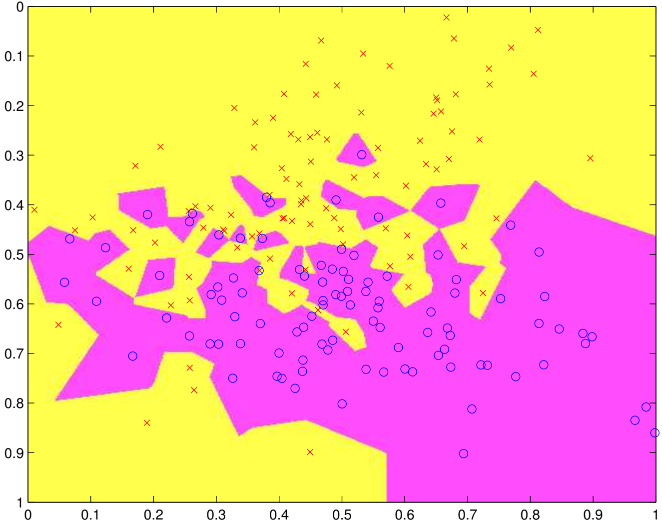
- Assume there is a “relevant” metric distance d on $\mathcal{X}$.
- General idea: if given a new point X, look at training points falling in the neighborhood of X.
- The 1-NN rule: predict for X the class of its nearest neighbor in the training set.
- The k-NN neighbor rule: same as above, but perform a majority vote among the k nearest neighbors.
- Weighted k-NN rule: same as above, with some weighted majority vote (weights according to the order of closeness)
- Note : this is a plug-in rule!

Example.



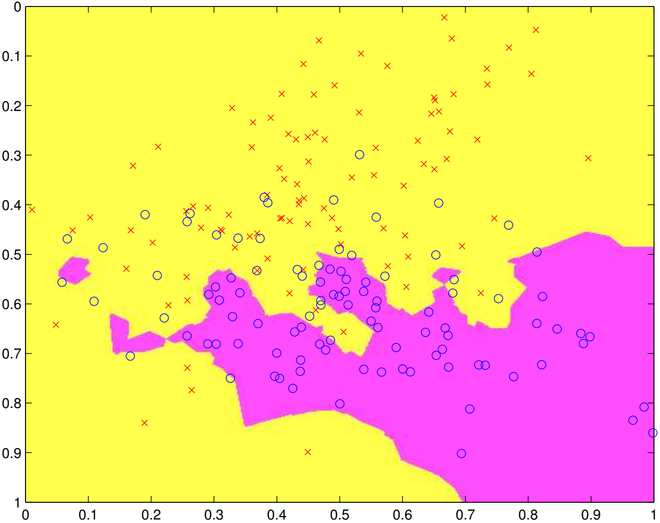
A classification problem (two Gaussians)

Example.



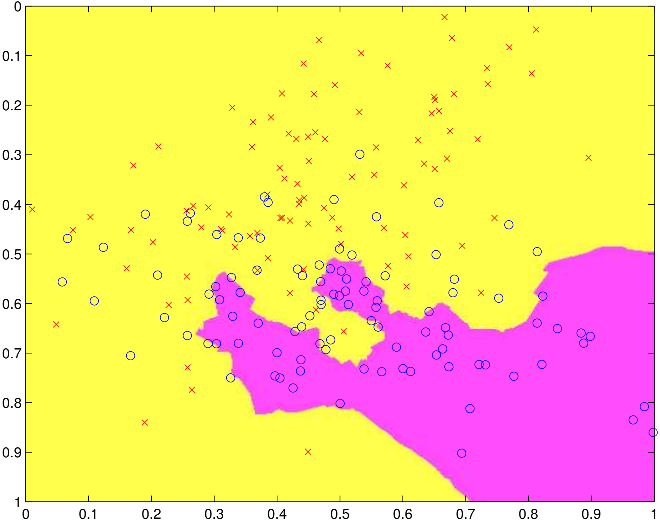
Result of 1-NN.

Example.



Result of 2-NN.

Example.



Result of 5-NN.

Goals for statistical analysis

- Note that unfortunately the NN procedure does not enter in a straightforward way in the framework of a classifier choice amound a fixed “set of functions”.
- We will be interested in the behavior of $\mathcal{E}(\widehat{f}_{k-NN})$ as the sample size grows to infinity:
  - For k fixed ?
  - As $k = k(n)$ varies with the sample size?

Asymptotics for k fixed: general idea

- We consider the binary clasification case.
- When the sample size is large, the k nearest neighbors $X^{(i)}(x)$ of x are very close to x.
- Then for the corresponding labels, $P(Y^{(i)} = 1|X = X^{(i)}) = \eta(X^{(i)})$ is very close to $P(Y = y|X = x) = \eta(x)$.
- If, for a fixed x, the $Y^{(i)}$ where, in fact, drawn according to $P(Y = y|X = x)$, then the average error conditional to $X = x$ would be

$$\mathcal{E}(\widehat{f}_{k-NN}(x)|X = x) = \eta(x)Q(\eta(x)) + (1 - \eta(x))(1 - Q(\eta(x))),$$

where $Q(u) = \mathbb{P}\left[Bin(u, k) > \lfloor \frac{k}{2} \rfloor\right]$.

First Lemma

Lemma

Let $X; X_1, X_2, \ldots$ *be drawn i.i.d.* $\sim P$. *Then*

$$d(X_n^{(k)}(X), X) \to 0$$

as $n \to \infty$, *in probability and a.s. (* $X^{(k)}$ *is the* k*-th NN among* $(X_i)_{1 \leq i \leq n}$ *.)*

Second Lemma

Lemma

Let L_k^* be the error of the "idealized" k-NN (where labels of neighbors of x would be drawn following $\eta(x)$), then

$$\mathbb{E}\left[\mathcal{E}(\widehat{f}_k) - L_k^*\right] \leq \sum_{i=1}^{k} \mathbb{E}\left[\left|\eta(X) - \eta(X^{(i)}(X))\right|\right] .$$

Principle: coupling argument.
If we assume $\mathcal{X}$ compact and η continuous , this leads to the conclusion.

Consequences of the asymptotics

- For c-classes problem, we can easily compute

$$L_1^* = 1 - \sum_{i=1}^{c} \mathbb{E}\left[P(Y = c|X)^2\right] .$$

- The asymptotic error L_k^* of k-NN is decreasing in k .
- We have the inequality

$$L^* \leq L_k^* \leq L^* + \sqrt{\frac{2L_1^*}{k}} \leq L^* + \sqrt{\frac{1}{k}} .$$

Behavior when L^* is small

Writing $L_k^* = \mathbb{E}\left[\alpha_k(\eta(x))\right]$, it is interesting to look at the behavior of $\alpha_k(p)$ as $p \to 0$:

- $\alpha_1(p) \sim 2p$;
- $\alpha_3(p) \sim p + 4p^2$;
- $\alpha_5(p) \sim p + 10p^3 \ldots$

If one assumes that L^* is "small", 3-NN is OK in an asymptotic sense (this is however not clear for a fixed sample size. . .)

Consistency

- Let $\widehat{f}^{(n)}$ be a sequence of classifiers for increasing sample size n.
- Weak consistency holds when

$$\mathbb{E}_{S_n}\left[\mathcal{E}(\widehat{f}^{(n)})\right] \to L^*$$

- Strong consistency holds if

$$\mathcal{E}(\widehat{f}^{(n)}) \to L^* \text{ a.s.}$$

- We have seen that the k-NN rule cannot be consistent (in general) if k is fixed. What if we allow k to depend on n?

Consistency of k-NN

Theorem

Assume $k(n) \to \infty$ and $k(n)/n \to 0$. Then the $k(n)$-NN rule is weakly consistent under either of the following conditions:
(i) $\mathcal{X}$ is compact and $\eta(x, y)$ is continuous;
(ii) $\mathcal{X} = \mathbb{R}^d$ and the distance is the standard euclidean one.

Note that in case (ii) we have (universal consistency), i.e., no assumptions have to be made at all on the generating distribution $P(X, Y)$.

Decomposition:

$$\mathcal{E}(\widehat{f}_{k(n)}^{(n)}) - L^* \leq 2\mathbb{E}\left[|\eta(X) - \widehat{\eta}_n(x)|\right]$$

(see first lecture)
Put

$$\overline{\eta}_n(x) = \frac{1}{k(n)} \sum_{i=1}^{k(n)} \eta(X^{(i)}(x))$$

then

$$\mathbb{E}\left[|\eta(X) - \widehat{\eta}_n(x)|\right] \leq \mathbb{E}\left[|\eta(X) - \overline{\eta}_n(x)|\right] + \mathbb{E}\left[|\overline{\eta}_n(X) - \widehat{\eta}_n(x)|\right]$$

Stone's Lemma

Lemma

Let f be any integrable function on $\mathbb{R}^d$. Then there exists a constant γ_d such that

$$\sum_{i=1}^{k} \mathbb{E}\left[\left|f(X^{(i)}(X))\right|\right] \leq k\gamma_d \mathbb{E}\left[|f(X)|\right].$$

- The universal consistency result, even for non-continuous η, might be counter-intuitive…
- Let's consider a particularly counter-intuitive example: learning to classify rational numbers from irrational ones!
- Assume:
  $P(Y = 1) = P(Y = 0) = 1/2$;
  $P(X|Y = 0) = \text{Uniform}([0, 1])$ (hence a.s. irrational);
  $P(X|Y = 1) =$ some discrete distribution on $\mathbb{Q}$
- …then the $k(n) - NN$ rule (following the requirements of the theorem) is consistent! Can we reconcile this with intuition?

Generalization of the consistency result:

Theorem

Consider a local averaging rule (for regression) of the form

$$\widehat{f}(x) = \sum_{i=1}^{n} W_{n,i}(x) Y_i \,,$$

where $(W_{n,i})$ *are a family of weights which may depend on the data. Then the following conditions are sufficient for consistency:*
(i) $\mathbb{E}\left[\sum_{i=1}^n W_{n,i}(X) f(X_i)\right] \leq c\mathbb{E}\left[f(X)\right]$ *for any nonnegative, integrable function f ;*
(ii) $\mathbb{E}\left[\max_i W_{n,i}(x)\right] \to 0$ *as* $n \to \infty$ *;*
(iii) for all $a > 0$, $\mathbb{E}\left[\sum_{i=1}^n W_{n,i} \mathbb{1}\{\|x - x_i\| \geq a\}\right] \to 0$ *as* $n \to \infty$ *;*

Implementation issues for k-NN

- The direct way to find a nearest neighbor takes $\mathcal{O}(n)$ operations.
- This can already be too expensive to compute if the training sample is large.
- Many methods exist either to obtain a faster computation or an approximate computation.
  - "prototype" methods: select a subset of the training set, of construct a reduced set of "prototypes" that are supposed to sum up the training set. Then apply k-NN using the prototype set.
  - "K-D trees": partition the data in a tree similar to a decision tree. Works when the dimension d is not too large.
  - In many cases the dimension is abitrary and/or the metric is non-euclidean. Then one must use only metric properties of the data.

Cover trees

- Cover trees are a recent method by Beygelzimmer, Kakade and Langford (2006) that yields extremely good results.
- A cover tree is a leveled tree structure where nodes where nodes are labeled by points. Denoting C_i the set of points at level i, the following properties hold:
  - (nesting) $\mathcal{C}_i \subset \mathcal{C}_{i-1}$.
  - (cover) Every $p \in \mathcal{C}_{i-1}$ has a parent $q \in C_i$ satifying $d(p,q) \leq 2^i$.
  - (separation) Any distinct points $q, q' \in \mathcal{C}_i$ satisfy $d(q,q') > 2^i$
- Cover Trees are a structure taking $\mathcal{O}(n)$ space where queries for the k-nearest neighbors of a new points are in $\mathcal{O}(\log n)$.

Side note: when is a distance matrix euclidean of dimension d?

- Assume we are given the $n \times n$ matrix of distances between any to points of a certain set.
- Can we ensure that these points can be represented in a d-dimensional eulidean space?
- Note: to construct a cover tree, in general it is not necessary to compute all the entries in the distance matrix.