Social Data Science: Machine Learning & Econometrics

Exercise class 1

February 19, 2020

Todays quick warmup

Q: The Collatz conjecture states that the sequence that arises from recursive application of

$$f(n) = \begin{cases} n/2, & n = 0 \pmod 2 \\ 3n+1, & n = 1 \pmod 2 \end{cases} \tag{1}$$

eventually cycles $4, 2, 1, 4, 2, 1, \ldots$ for any $n \in \mathbb{N}$. Implement a *recursive* version of $f(n)$ and check if the conjecture holds for $n = 10$, $n = 11$, and $n = 987654321$.

Hint: What should `collatz(n)` return if $n = 2$? What if n is even? And finally what if n is odd?

Todays quick warmup - solution

Pretty self-explanatory, % is for modulo.

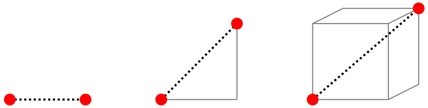
```
def collatz(n):
    if n == 2: return 1
    if n % 2 == 0:
        return collatz(n/2)
    if n % 2 == 1:
        return collatz(3*n + 1)
```

Last lecture in a nutshell

Last lecture covered a lot

- Decision Trees and random forests.
- KNN and kernel methods.
- Bagging and boosting

KNN simply predicts the average (or majority) of a points k neighbors.
The weakness of KNN is the curse of dimensionality:



Last lecture in a nutshell

Bagging means bootstrapping predictions; a model which is bagged over B bootstrap samples computes

$$\hat{f}_b(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x)$$

Bagging improves fit by reducing sensitivity to instability in the model procedure.

Boosting fits and sums simple base learners on "residualized" data. With squared-error-loss and continuous labels boosting is exactly repeated regression on residuals. For different loss-functions use the pesudo-residuals

$$r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}}.$$