Uncertainties in blood flow calculations and data

Rachael Brag and Pierre Gremaud (NCSU)

August 10, 2014

Goals

1. introduce methods from machine learning
   - *Machine learning (...) deals with the construction and study of systems that can learn from data, rather than follow only explicitly programmed instructions.* Wikipedia
   - *Machine learning is the science of getting computers to act without being explicitly programmed.* E. Ng

$$x \longrightarrow \boxed{\text{nature}} \longrightarrow y$$

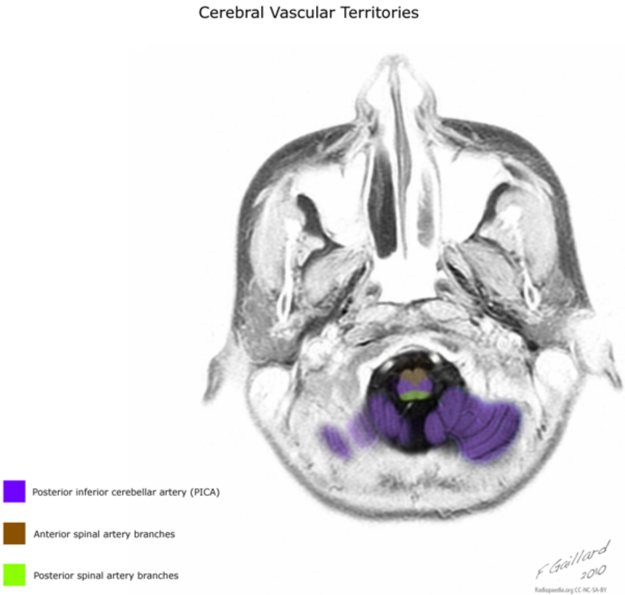
math, stat: $x \longrightarrow \boxed{\text{linear regression, other models}} \longrightarrow y$

machine learning: $x \longrightarrow \boxed{\phantom{\text{nature}}} \longrightarrow y$

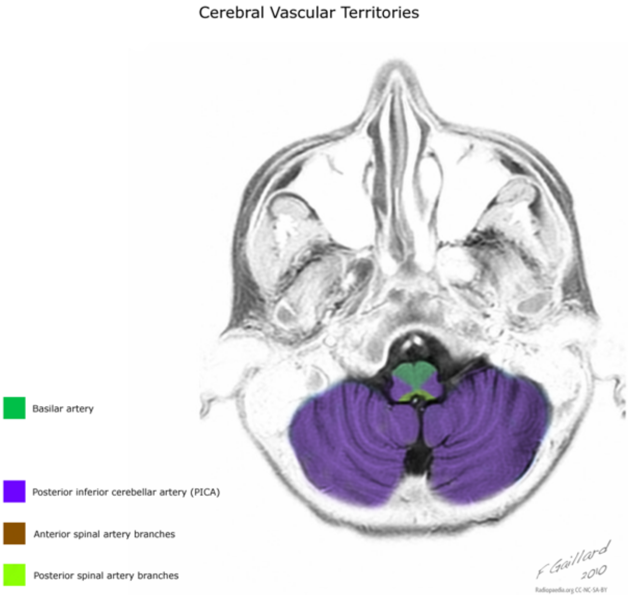
$\searrow$ regression trees $\nearrow$

2. illustrate new concepts on Cerebral Blood Flow (CBF) studies

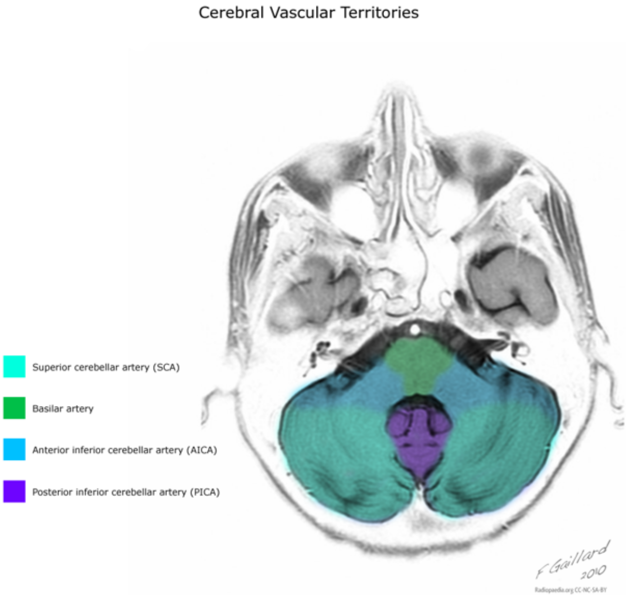
Vascular territories



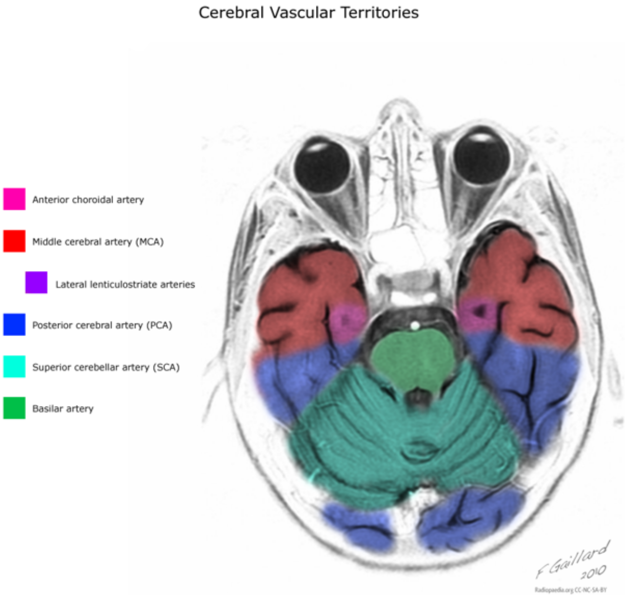
Vascular territories



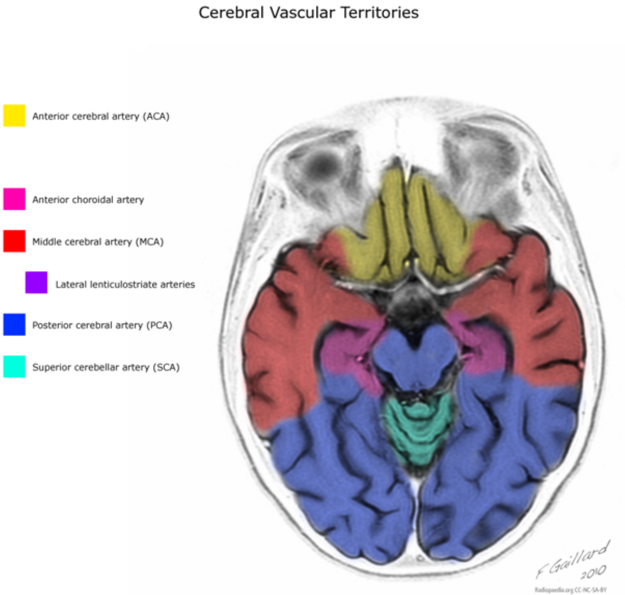
Vascular territories



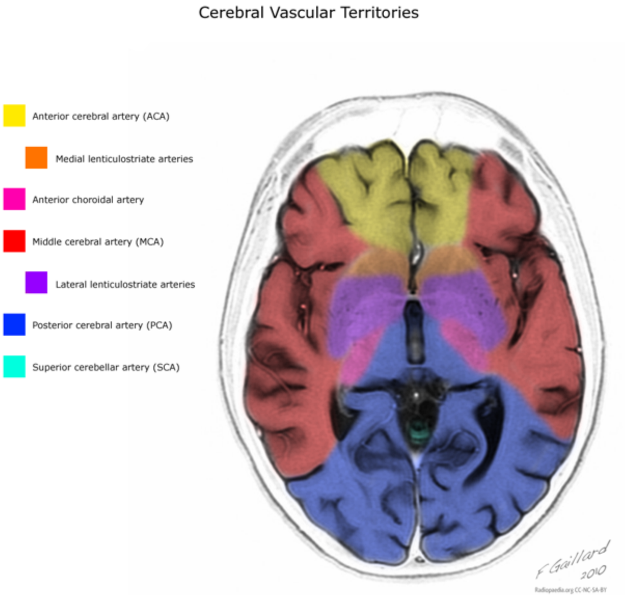
Vascular territories



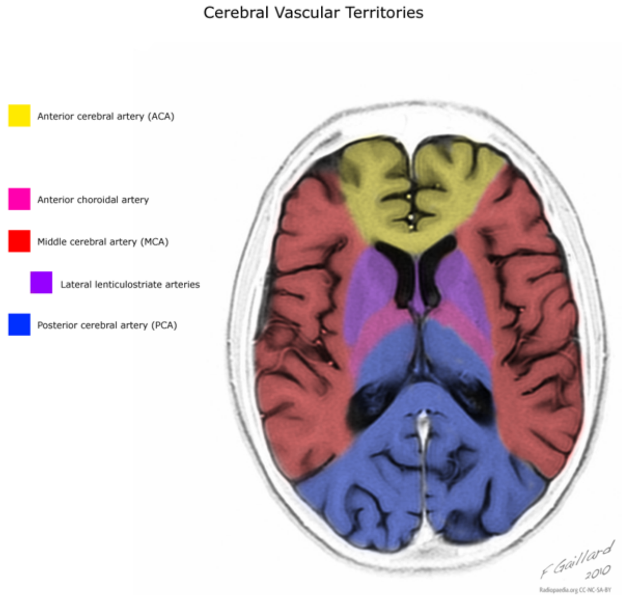
Vascular territories



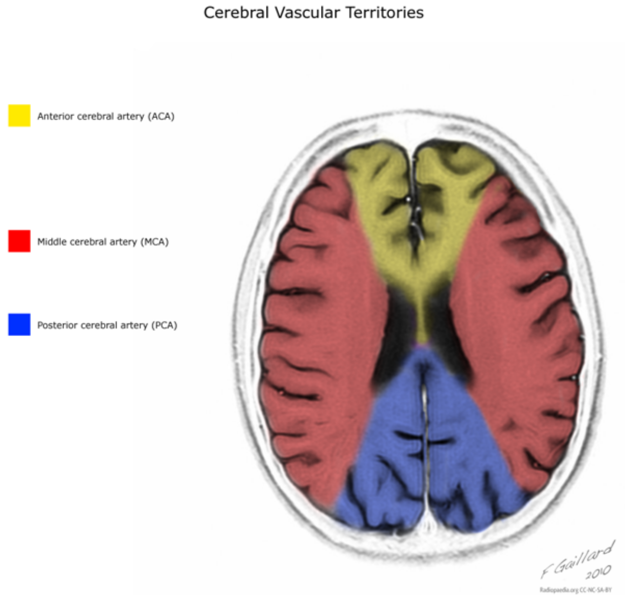
Vascular territories



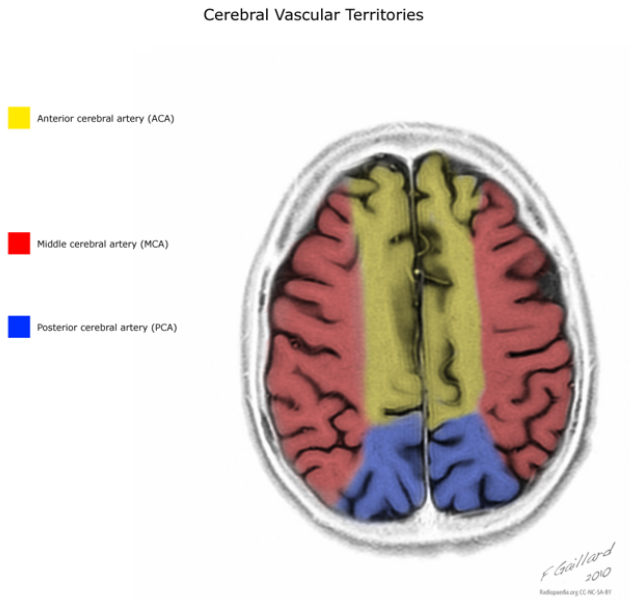
Vascular territories



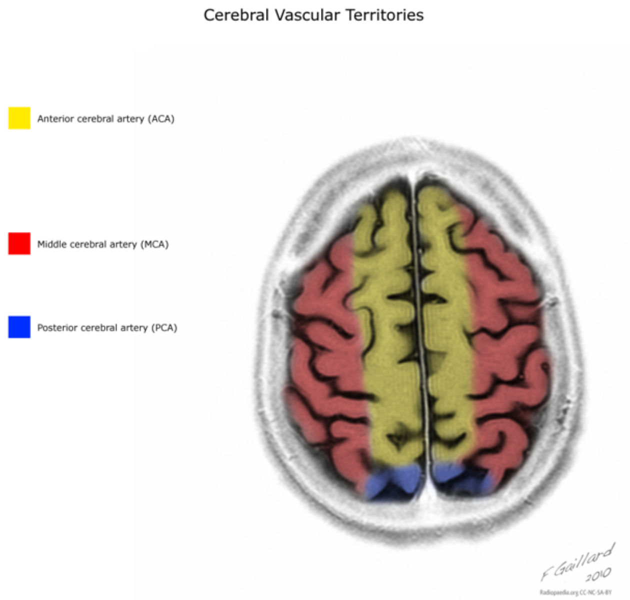
Vascular territories



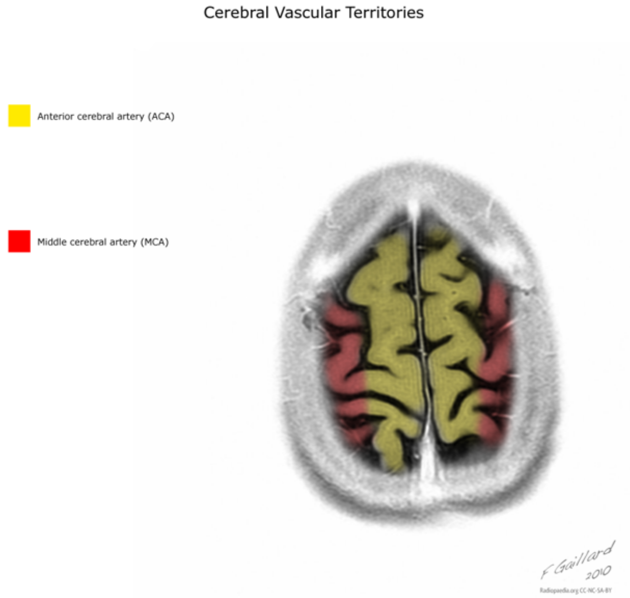
Vascular territories



Vascular territories



Vascular territories



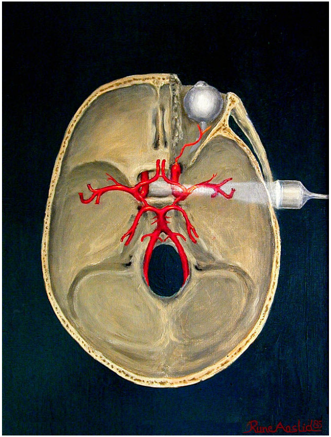
Problem

Can we estimate local CBF

- cheaply
- continuously and in real time
- accurately
- or at least with "error bars"?

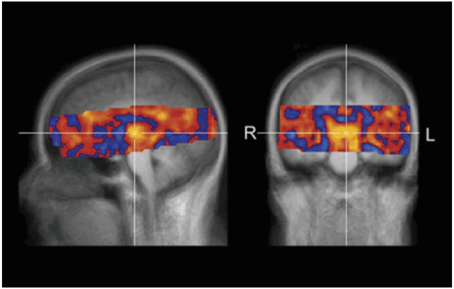
Problem

cheap: Transcranial Doppler (TCD)



wikipedia

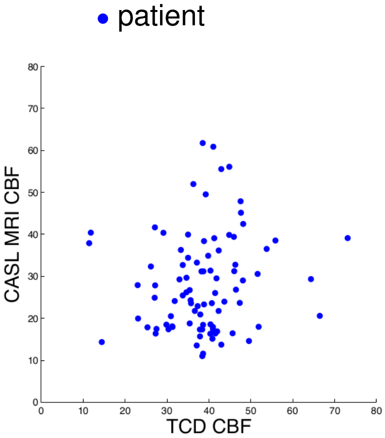
expensive: Magnetic Resonance Imaging (MRI)



Mangia et al., J. Cereb. Blood Flow Metab., 32 (2012)

Problem

This $\longrightarrow$



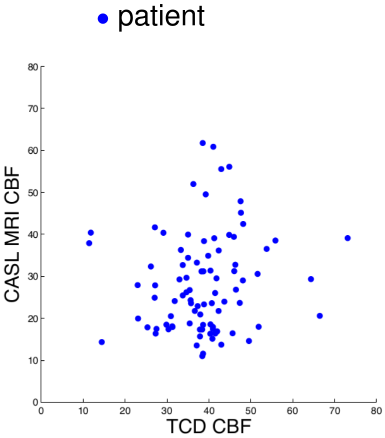
$\uparrow$

"should" agree with that

Problem

This ⟶

Oy!



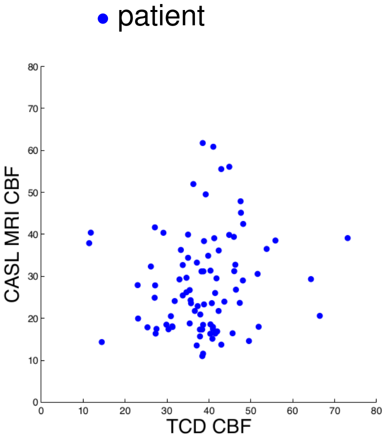
↑

"should" agree with that

Problem

This ⟶

"It is intriguing that methods measuring the same physiological parameter do not correlate." *Henriksen et al. J. Magn. Res. Imaging, 2012*



↑

"should" agree with that

Hypothesis

different patients react differently to the measurement protocols

so...

- let's group patients into "like" groups
- let's apply local "models" in each group

to do so, we let the "data speak"....

Overview

- linear and nonlinear approximations
- local regression and trees
- classification
- random forests
- back to CBF, UQ and other acronyms

Mathematical challenge

- predictor variable (vector): $x = [x_1, \dots, x_d]$
- response variable (scalar): y

WANTED: value (or distribution) of y for given x, i.e.

$$y = f(x)$$

CHALLENGE: we do not have f but "just" data

$$[x_i, y_i] = [x_{i,1}, \dots, x_{i,d}, y_i], \qquad i = 1, \dots, N$$

For us: $d = 14$, $N =$ number of patients ≈ 200

Approximation 101: linear

"Pretend" we know f and $x \in \Omega = [0,1]^d$

- partition Δ of Ω into cells ω
- piecewise constant (to simplify) approximation

$$f_h(x) = \sum_{\omega \in \Delta} c_\omega \chi_\omega(x)$$

- best constants: $c_\omega = \frac{1}{|\omega|} \int_\omega f(x)\,dx$ = mean of f on ω
- well know result:

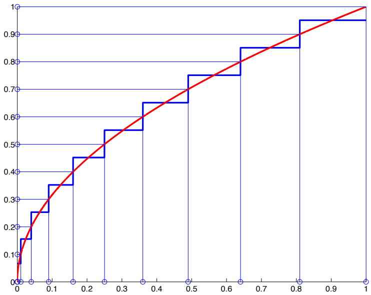
$$\|f - f_h\| \leq C(d) N^{-1/d} \|\nabla f\|$$

$N = m^d$ = number of cubes of length $h = 1/m$

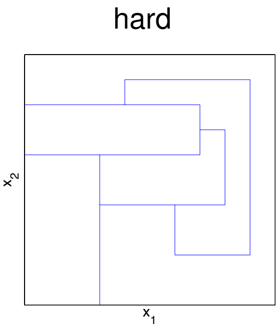
Approximation 102: nonlinear

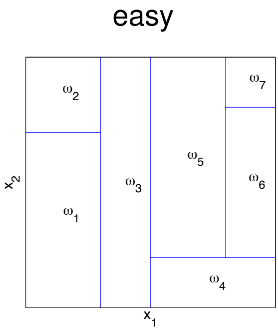
Choose better partitions based on f/data

- "equivariation" partition (Kahane 1961)
- easy in 1d (partition depends on f)
- "optimal" partitions in higher dim not doable

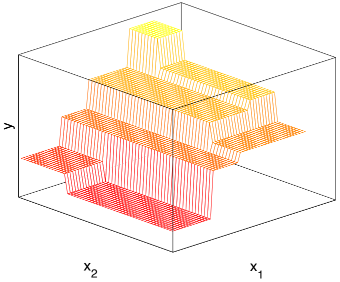


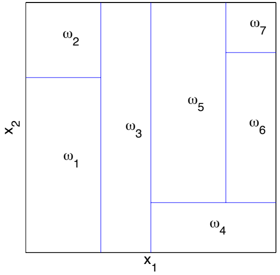
Minimization $\longrightarrow$ recursive partitioning



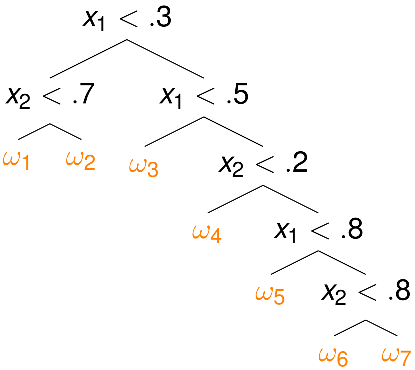


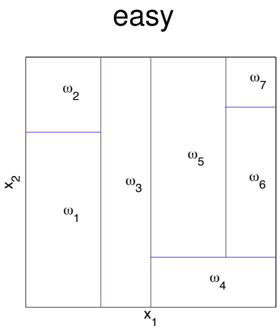
Minimization $\longrightarrow$ recursive dyadic partitioning





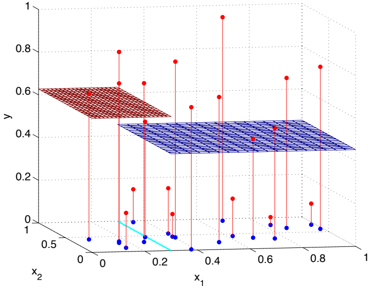
Minimization $\longrightarrow$ recursive dyadic partitioning

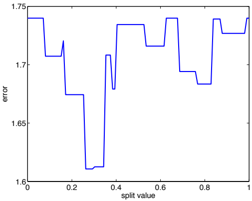




Trees and data

- loop on split variables x_j, $j = 1, 2, \ldots$
  - loop on split split values s
    - $\omega_1(j,s) = \{x; x_j \le s\}$, $\omega_2(j,s) = \{x; x_j > s\}$
    - error = $\min_{j,s}\left\{\sum_{x_i \in \omega_1(j,s)} (y_i - c_1)^2 + \sum_{x_i \in \omega_2(j,s)} (y_i - c_2)^2\right\}$
  - end
- end





baby tree

$x_1 < .3$

$\omega_1 \quad \omega_2$

Regression tree

1. consider all binary splits on every predictor
2. select split with lowest MSE and $|\text{child node}| < \mathit{MinLeaf}$
3. impose split
4. repeat recursively for child nodes

Stop if any of the following holds

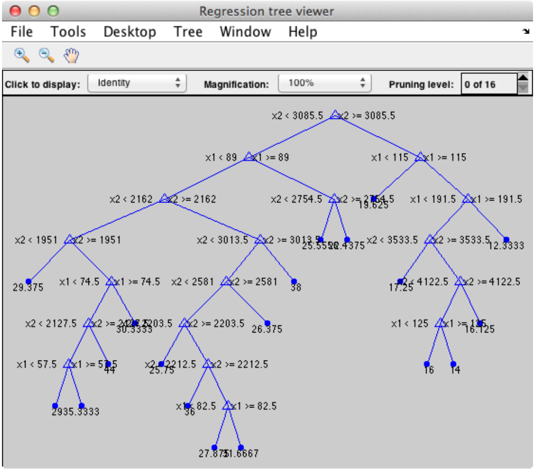
- node is pure (MSE $<$ *qetoler* $\times$ MSE(full data))
- fewer than *MinParent* observations in node
- $|\text{child node}| < \mathit{MinLeaf}$

MATLAB example

» LOAD CARSMALL

» X = [HORSEPOWER WEIGHT];

» RTREE = FITRTREE(X,MPG);



Classification tree

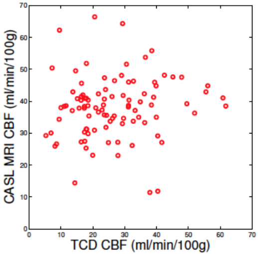
- What about categorical variables? (gender (F/M), diabetes (Y/N), hypertensive (Y/N), car manufacturer (AMC/Aston Martin/Ferrari/Datsun/Peugeot/Rolls Royce/Yugo etc...)
- MSE $\longrightarrow$ Gini impurity

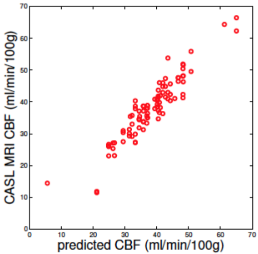
$$\sum_{k=1}^{K} p_{mk}(1 - p_{mk})$$

  - $p_{mk} = \frac{1}{|\omega_m|} \sum_{x_i \in \omega_m} \delta_{x_i,k}$ = fraction of items from class k in ω_m
  - how often a randomly chosen element from ω_m would be incorrectly labeled if it were randomly labeled according to the distribution of classes in ω_m
  - issues with mixed data...

Does this stuff work?

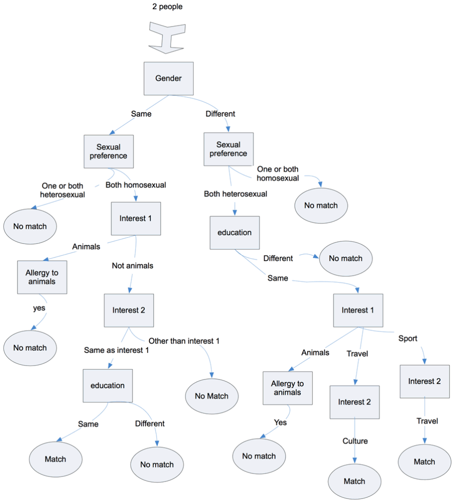
Yes! MSE divided by ≈ 4





What's good about trees

1. easy to understand



What's good about trees

1. easy to understand
2. can handle both categorical and numerical predictors
3. can handle missing data
4. fast
5. no model!

What's not so good about trees

1. trees are unstable
2. predictions are not smooth
3. biases toward predictor variables with high variation
4. no model $\Rightarrow$ little analysis

Doing better: bagging

bootstrap aggregating

- for $b = 1$ to B
  - draw bootstrap sample of size N from training data (uniformly and with replacements)
  - grow tree T_b to bootstrapped data
- end
- average to get prediction for x:

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

Issues with bagging

- trees T_b's are correlated: i.d. but not i.i.d.
- i.i.d: $\mathrm{var}(\sum_i X_i) = \sum_i \mathrm{var}(X_i) \Rightarrow$

$$\mathrm{var}(\hat{f}(x)) = \frac{\sigma^2}{B}$$

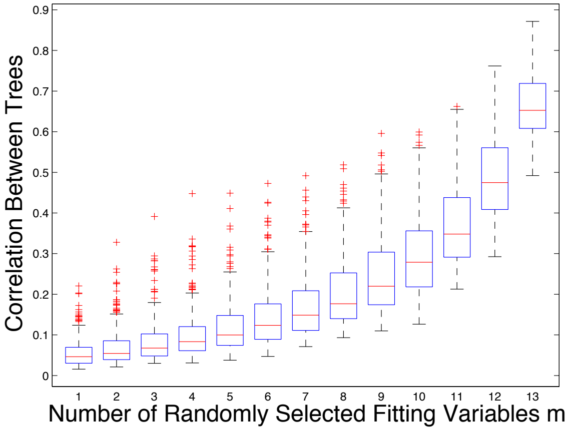
- correlated i.d:
$\mathrm{var}(\sum_i X_i) = \sum_i \mathrm{var}(X_i) + 2\sum_{i<j} \mathrm{cov}(X_i, X_j) \Rightarrow$

$$\mathrm{var}(\hat{f}(x)) = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$$

- $\rho \downarrow$ and $B \uparrow \Rightarrow$ variance $\downarrow$

Random forests (Breiman 2001)

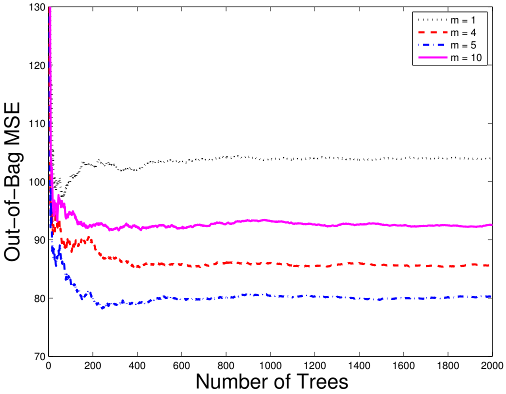
decrease tree correlation by splitting based on $m < d$ variables



OOB errors

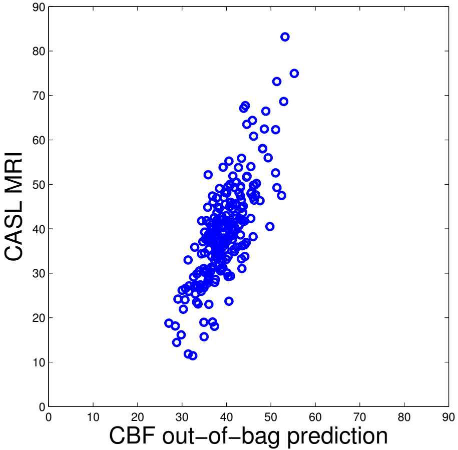
error check on training data

- for each (x_i, y_i), construct RF predictor by averaging only trees from bootstrap samples not containing (x_i, y_i)



$m = 5 < d = 14$ wins

Results for our problem



MSE divided by ≈ 8

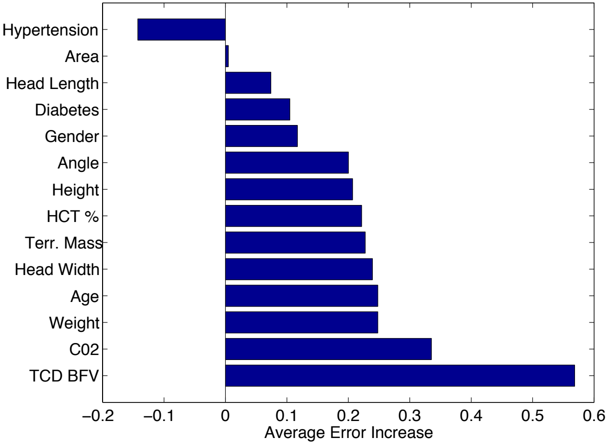
But wait, there is more...

Trees can be used to assess variable importance

1. Gini importance: at each split, MSE reduction attributed to split variable and accumulated over all trees for each variable $\Rightarrow$ bias toward high variability predictors
2. permutation importance: in each tree, compute MSE for OOB samples; then randomly sample values of variable and compute increase in OOB MSE

room for improvements and analysis...

Variable importance

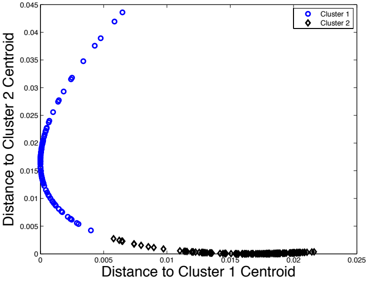


Clustering

- consider each pair of patients
- count number of times pair belongs to same tree in the forest

$\Rightarrow$ proximity matrix A

- clustering algorithms (spectral or other) can be applied to A



Conclusion

- machine learning: powerful for "messy" problems
- simple, efficient
- may be hard to interpret and analyze
- low hanging fruits for mathematicians...

More references

literature

Statistical modeling: the two cultures, L. Breiman, Statistical Sc., 16 (2001), p. 199–231.

The elements of statistical learning, T. Hastie, R. Tibshirani, J. Friedman, Second Edition, Springer Series in Statistics, 2009

Cerebral blood flow measurements: ..., R. Bragg, P.A. Gremaud, V. Novak, in preparation

software

MATLAB FITENSEMBLE from the stat toolbox

R RANDOMFOREST package

java `http://www.cs.waikato.ac.nz/ml/weka/`