Alternative form of bagging

This suggests an alternative method for bagging. Now given an input $x \in \mathbb{R}^p$, instead of simply taking the prediction $\hat{f}_{\text{tree}, b}^{\text{tree}}(x)$ from each tree, we go further and look at its predicted class probabilities $\hat{p}_{k}^{\text{tree}, b}(x)$, $k = 1, \ldots, K$. We then define the bagging estimates of class probabilities:

$$\hat{p}_{k}^{\text{bag}}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{p}_{k}^{\text{tree}, b}(x) \quad k = 1, \ldots, K$$

The final bagged classifier just chooses the class with the highest probability:

$$\hat{f}^{\text{bag}}(x) = \arg\max_{k=1, \ldots, K} \hat{p}_{k}^{\text{bag}}(x)$$

This form of bagging is preferred if it is desired to get estimates of the class probabilities. Also, it can sometimes help the overall prediction accuracy...

...From Ryan Tibshirani
Ocean range classification

16*17*2/2*60 = 16000 features

100 output classes (ranges)

120 training cases (one ship track)

How is boosting performed?

FIG. 11. (Color online) Source localization as a classification problem. Range predictions on Test-Data-1 (a, b, c) and Test-Data-2 (d, e, f) by FNN, SVM and RF for 300–950 Hz with 10 Hz increment, i.e., 66 frequencies. (a), (d) FNN classifier, (b), (e) SVM classifier, (c), (f) RF classifier.
Bagging example (Hastie)

DATA: One Tree

- Deterministic problem; noise comes from sampling distribution of $X$.
- Use a training sample of size 200.
- Here *Bayes Error* is 0%.
Bagging example (Hastie)

**Decision Boundary: Bagging**

Error Rate: 0.032

- Bagging averages many trees, and produces *smoother* decision boundaries.
- Bagging reduces error by variance reduction.
- Bias is slightly increased because bagged trees are slightly shallower.
PATTERN RECOGNITION AND MACHINE LEARNING

CHAPTER 8: GRAPHICAL MODELS
Bayesian Curve Fitting (1)

Polynomial

\[ y(x, \mathbf{w}) = \sum_{j=0}^{M} w_j x^j \]

\[ p(t, \mathbf{w}) = p(\mathbf{w}) \prod_{n=1}^{N} p(t_n | y(\mathbf{w}, x_n)) \]

[Diagram of Bayesian network with plates]

Plate
Define a conjugate prior over $w$

$$p(w) = \mathcal{N}(w|m_0, S_0).$$

Combining this with the likelihood function and using results for multiplying Gaussians, gives the posterior

$$p(w|t) = \mathcal{N}(w|m_N, S_N)$$

$$m_N = S_N \left( S_0^{-1} m_0 + \beta \Phi^T t \right)$$

$$S_N^{-1} = S_0^{-1} + \beta \Phi^T \Phi.$$

$$p(w, t) = p(t|w) p(w)$$

A common simpler prior

$$p(w) = \mathcal{N}(w|0, \alpha^{-1} I)$$

Which gives

$$m_N = \beta S_N \Phi^T t$$

$$S_N^{-1} = \alpha I + \beta \Phi^T \Phi.$$
Lecture 5: Predictive Distribution

Predict \( t \) for new values of \( x \) by integrating over \( w \) (Giving the marginal distribution of \( t \)):

\[
p(t|t, \alpha, \beta) = \int p(t|w, \beta) p(w|t, \alpha, \beta) \, dw
\]

\[
= \mathcal{N}(t|\mathbf{m}_N^T \phi(x), \sigma^2_N(x))
\]

where

\[
\mathbf{m}_N = \beta \mathbf{S}_N \Phi^T \mathbf{t}
\]

\[
\mathbf{S}_N^{-1} = \alpha \mathbf{I} + \beta \Phi^T \Phi.
\]

\[
\sigma^2_N(x) = \frac{1}{\beta} + \phi(x)^T \mathbf{S}_N \phi(x).
\]

\[
p(t) = \int p(t, w) \, dw = \int p(t|w) p(w) \, dw
\]
Predictive distribution: \[ p(\hat{t}|\hat{x}, x, t, \alpha, \sigma^2) \propto \int p(\hat{t}, t, w|\hat{x}, x, \alpha, \sigma^2) \, dw \]

where

\[ p(\hat{t}, t, w|\hat{x}, x, \alpha, \sigma^2) = \prod_{n=1}^{N} p(t_n|x_n, w, \sigma^2) \]

\[ p(w|\alpha)p(\hat{t}|\hat{x}, w, \sigma^2) \]
Conditional Independence

$a$ is independent of $b$ given $c$

\[ p(a | b, c) = p(a | c) \]

Equivalently

\[
\begin{align*}
p(a, b | c) &= p(a | b, c) p(b | c) \\
&= p(a | c) p(b | c)
\end{align*}
\]

Notation

\[ a \perp b \mid c \]

\[ a \& b \text{ independent} \]

\[ p(a, b) = p(a) p(b) \]
Conditional Independence: Example 1

\[
P(a, b, c) = p(c) \cdot p(a | c) \cdot p(b | c) \\
P(a, b) = \int p(c) \cdot p(a | c) \cdot p(b | c) \, dc
\]
\[
a \perp b \mid \emptyset
\]

\[
p(a, b, c) = p(a, b | c) \cdot p(c)
\]
\[
\frac{p(a, b, c)}{p(c)} = p(a, b | c) = p(a | c) \cdot p(b | c)
\]
\[
a \perp b \mid c
\]
Conditional Independence: Example 1

\[ p(a, b, c) = p(a|c)p(b|c)p(c) \]

\[ p(a, b) = \sum_c p(a|c)p(b|c)p(c) \]

\[ a \perp b \mid \emptyset \]

\[ p(a, b|c) = \frac{p(a, b, c)}{p(c)} = p(a|c)p(b|c) \]

\[ a \perp b \mid c \]
Conditional Independence: Example 2

\[
p(a, b, c) = p(a) p(c | a) p(b | c)
\]
\[
p(a, b) = \int p(a) \int p(c | a) p(b | c) \, dc
\]

\[
a \perp \! \! \! \perp b \mid \emptyset
\]

\[
\frac{p(a, b, c)}{p(c)} = \frac{p(a | c) p(c) p(b | c)}{p(c)} = p(a | c) p(b | c)
\]

\[
a \perp \! \! \! \perp b \mid c
\]
Conditional Independence: Example 2

\[ p(a, b, c) = p(a)p(c|a)p(b|c) \]

\[ p(a, b) = p(a) \sum_c p(c|a)p(b|c) = p(a)p(b|a) \]

\[ a \perp\!
\perp b \mid \emptyset \]

\[ p(a, b|c) = \frac{p(a, b, c)}{p(c)} = \frac{p(a)p(c|a)p(b|c)}{p(c)} = p(a|c)p(b|c) \]
Conditional Independence: Example 3

\[ p(a, b, c) = p(a) p(b) p(c | a, b) \]
\[ p(a, b) = p(a) p(b) \]
\[ a \perp b \mid \emptyset \]

\[ p(a, b, c) = \frac{p(a)p(b)}{p(c)} p(c | a, b) \]
\[ a \not\perp b \mid c \]

Note: this is the opposite of Example 1, with c unobserved.
Conditional Independence: Example 3

\[ p(a, b, c) = p(a)p(b)p(c|a, b) \]

\[ p(a, b) = p(a)p(b) \]

\[ a \perp b \mid \emptyset \]

\[ p(a, b|c) = \frac{p(a, b, c)}{p(c)} = \frac{p(a)p(b)p(c|a, b)}{p(c)} \]

\[ a \not\perp b \mid c \]

Note: this is the opposite of Example 1, with c unobserved.
\[ p(x | y) = \frac{p(x)p(y | x)}{p(y)} \]

\[
p(y) = \sum_{x'} p(y | x') p(x')
\]

\[
p(x | y) = \frac{p(y | x)p(x)}{p(y)}
\]
\[ p(\{x_m, \mu_m\}) = p(x_1 | \mu_1) p(\mu_1) \prod_{m=2}^{M} p(x_m | x_{m-1}, \mu_m) p(\mu_m) \]

\[ p(\mu_m) = \text{Dir}(\mu_m | \alpha_m) \]

\[ O(k^M) \]
State space model

Measurement

\[ y_k = H_k x_k + v_k \]

\[ v_k \sim N(0, R_k) \]

State Eq

\[ x_{k+1} = F_k x_k + w_k \]

\[ w_k \sim N(0, Q_k) \]
State space model

\[
X_0 \sim N(0, P_0)
\]

State Eq.
\[
X_{k+1} = M_k X_k + S_k
\]

Measurement Eq.
\[
y_k = H_k X_k + V_k
\]

Predict:
\[
X_{k|k-1} \quad \text{Update} \quad X_k
\]

\[
S_k \sim N(0, Q_k)
\]

\[
V_k \sim N(0, R_k)
\]
Graph clustering for localization within a sensor array

Peter Gerstoft and Nima Riahi,  noiselab.ucsd.edu
Christoph Mecklenbrauker, TU Wien

Based on paper: Riahi and Gerstoft, Signal Processing, 2017

March 5—12, 2011: 3TB, 5200 Stations in Long Beach, California
Location 1: Prince - “Sign o’ the times”

$f = 750 \text{ Hz}$

Location 1: Otis Redding - “Hard to handle”

Spectral coherence between $i$ and $j$

$\hat{C}_{ij}(f) = \frac{1}{N} \sum_{t=1}^{N} X_i(f,t) \cdot \bar{X}_j(f,t)$

(Normalization: $|X(f,t)|^2 = 1$)
Statistically significant entries => Connectivity matrix

Each group is spatially coherent. But no temporal correlation between groups (i.e. different source)

- Each sensor is a node in the graph.
- If nodes $i$ and $j$ are significantly correlated $|C_{ij}| > \xi$, then they share an edge.
Two sources in the network

Statistically significant entries => Connectivity matrix

Graph with 30 nodes

Connected subgraphs:
- 5 nodes and 9 edges
- 8 nodes and 20 edges

- Each sensor is a node in the graph.
- If nodes $i$ and $j$ are significantly correlated $|C_{ij}| > \xi$, then they share an edge.
- A subgraph has high spatial coherence across part of the array (and likely a source nearby).
Asymptotic case

Reinterpret $C_{ij}$ as connectivity matrix $E_{ij}$ of network with $N$ vertices.

$$E_{ij}^0 = \begin{cases} 1 & \text{if } \hat{C}_{ij} > c_\alpha \\ 0 & \text{otherwise}, \end{cases}$$
If $\alpha > 2.5/(N-1)$ the array network almost surely has a giant connected component, i.e. almost all sensors are indirectly linked to each other \cite{Erdos1959}.

**Bad for cluster search!**
"Noise-only" network
If $\alpha > 2.5/(N-1)$ the network almost surely has a giant connected component, i.e., most sensors are linked [Erdös & Rényi, 1959].

**Bad for cluster search!**

We limit this by testing just the 8-nearest neighbors:

$$E_{ij} = \begin{cases} 1 & \text{if } \hat{C}_{ij} > c_{\alpha} \text{ and } i \in N(j) \\ 0 & \text{otherwise,} \end{cases}$$

**Connectivity matrix => Band structure**

Prohibit long-range coherences
Helicopter rotor noise (seismo-acoustic coupling)

Several peaks consistent with helicopter rotor harmonics (20-100 Hz).

Doppler shift

\[ \frac{f_{\text{high}}}{f_{\text{low}}} = \frac{(v_0 + v)}{(v_0 - v)} \approx 1.4 \text{ i.e. } v \approx 250 \text{ km/h} \]

Speed over ground 7km/2min=210km/h

✓ Rotor frequencies
✓ Doppler frequency shift
✓ Movement in map
Clusters on March 10

Based on 9400 time windows x 10 frequency bins.
Each dot is the center of a cluster. 90% of the clusters cover <1.5% of the area.
Few false detections