Lecture 9: Testing Juntas

Lecturer: Ronitt Rubinfeld
6.893 Spring 2013
Junta functions

Def. $f$ is a $k$-junta if it depends on $\leq k$ variables.
Juntas

- How many relevant features in the data?
- Interesting also for non-Boolean functions
- Can we test whether a function is a Junta?
A first idea:

- Find all relevant variables
- If more than k, reject
Find *any* relevant variable:

- Find any \( x, y \) such that \( f(x) \neq f(y) \)
- \( S \leftarrow \{i \mid x_i \neq y_i\} \) must contain a relevant variable
“Hybrid” argument

\[ x = x^{(0)} = x_1, x_2, \ldots, x_n \]
\[ x^{(1)} = y_1, x_2, \ldots, x_n \]
\[ x^{(2)} = y_1, y_2, y_3, \ldots, x_n \]
\[ \ldots \]
\[ x^{(n)} = y_1, \ldots, y_n \]

\[ f(x) = f(x^{(0)}) \]
\[ f(x^{(1)}) \]
\[ f(x^{(2)}) \]
\[ \ldots \]
\[ f(x^{(n)}) = f(y) \]

Since \( f(x) \neq f(y) \) there must be at least one \( i \) such that
\[ f(x^{(i)}) \neq f(x^{(i+1)}) \]

Any such \( i+1 \) is relevant!
How many queries?

- All $n$ hybrids?
- $O(\log n)$ queries: Can find $i$ via binary search
  - if $f(x^{(0)}) \neq f(x^{\left(\frac{n}{2}\right)})$ recurse on $0..n/2$
  - else recurse on $n/2+1..n$

- Better than $O(\log n)$?
  - don’t need to find $i$
Solve a weaker problem

- Partition variables into “groups”
- Find all groups containing relevant variables in $O(\log \# groups)$ queries
- If found $>k$ groups, then not a $k$-junta

Idea: partition randomly
Algorithm (given $k, \epsilon$)

- randomly partition $1..n$ into $s = poly(k, 1/\epsilon)$ parts $I_1, \ldots, I_s$
- $R \leftarrow \emptyset$
- Repeat up to $r = O\left(\frac{k}{\epsilon}\right)$ times:
  - generate $x, y$ randomly such that $x_R = y_R$
  - If $f(x) \neq f(y)$ use binary search to find relevant $I_i$
  - $R \leftarrow R \cup I_j$
  - if $R$ has $>k$ relevant parts, REJECT

agree on indices in “relevant” parts

Main issue for analysis: How long to find such an $x, y$?
Progress?

- if $f(x) \neq f(y)$ then must have found a new relevant partition, since all variables in R were set identically
Analysis

- If $f$ is a $k$-junta, always accepts
- If $f$ is $\epsilon$-far....
  - need to show that with high probability over choice partition $I_1, \ldots, I_s$ and for all $R$ containing $\leq k$ parts, $\Pr_{x,y \text{ s.t. } x_R = y_R} [f(x) \neq f(y)]$ is high
Notation

- $X_S$ is an ordered list, i.e. $(x_i : i \in S)$

- $X_S Y_{\bar{S}} \equiv Z = (z_1 \ldots z_n)$ such that $Z_S = X_S$ and $Z_{\bar{S}} = Y_{\bar{S}}$
Definition of Influence

- given $f$, Influence of $S \subseteq [n]$ is

$$\text{Inf}_f(S) = 2 \Pr_{x, y \text{ s.t. } x \bar{S} = y \bar{S}} [f(x) \neq f(y)]$$

- Factor of 2 allows matching standard single variable definition
- equivalent to $2 \Pr_{x, y} [f(x) \neq f(y_{S}x \bar{S})]$
- Note that $1/\text{Inf}_f(S)$ directly related to time to find pair $x, y$ s.t. $f(x) \neq f(y)$
Useful Proposition

- $\text{Inf}_f(S) = \sum_{T: S \cap T \neq \emptyset} \hat{f}(T)^2 = \sum_{T \subseteq [n]} \hat{f}(T)^2 - \sum_{T \subseteq \bar{S}} \hat{f}(T)^2$

- Why? similar reasoning to last lecture. (see also handwritten notes)

- Useful corollary: (monotonicity, subadditivity)

\[
\text{Inf}_f(S) \leq \text{Inf}_f(S \cup S') \leq \text{Inf}_f(S) + \text{Inf}_f(S')
\]

(why? by looking at terms in sums of squares of Fourier coefficients).
Important Lemma

Lemma: $f$ is $\varepsilon$-far for $k$-junta $\Rightarrow$

$\forall J \text{ s.t. } |J| \leq k, \quad \text{Inf}([n] \setminus J) \geq \varepsilon$

So, fixing any $k$ vars, leaves $f$ with influence $\geq \varepsilon$

Why useful?

As algorithm progresses (fixing more vars), still have enough influence to make progress!
One idea from the proof:

- Fix $J$ s.t. $|J| \leq k$
- define $h$ s.t. $h(x) = \text{sign}(E_z[f(x_Jz_j)])$

Note:
- $h$ is \textit{best} junta fctn on $J$

Rest of the proof: Fourier + other analysis of $h$...
Lemma is not enough!

- Need to redo lemma for fixing any $k$ PARTS
  - that’s a lot of variables!
Towards a main Lemma

Def. $f$ is a $k$-part junta wrt $\mathcal{I}$ if relevant coordinates in $\leq k$ parts of $\mathcal{I}$

Def. $f$ is $\epsilon/2$-far from $k$-part junta wrt $\mathcal{I}$ if $\forall J$ s.t. $J$ is union of $k$ parts in $\mathcal{I}$:

$$\inf_{f}([n]\setminus J) \geq \epsilon/2$$

(note: not standard definition of “far” – stronger?)
Main Lemma

Given $f$ that is $\epsilon$-far from a k-junta.

Let $\mathcal{X}$ be a random partition with $s = poly(k, \frac{1}{\epsilon})$ parts.

With probability $\geq \frac{5}{6}$,

$$f \text{ is } \epsilon/2\text{-far from a k-part junta w.r.t. } \mathcal{X}$$

(i.e., $\text{Inf}_f([n] \setminus J) = \sum_{S \subseteq [n]} \hat{f}(S)^2 - \sum_{S \subseteq J} \hat{f}(S)^2 \geq \epsilon/2$)
Theorem: There is an Algorithm $T$ using $O\left(\frac{k}{\epsilon} + k \log k\right)$ queries, satisfying:

- If $f$ is a $k$-junta, $T$ outputs PASS
- If $f$ is $\epsilon$-far from a $k$-junta, $\Pr[T$ outputs fail$]>2/3$
Main Lemma $\rightarrow$ Theorem

- If $f$ is a $k$-junta, $T$ finds at most $k$ relevant parts and thus accepts
- If $f$ is $\epsilon$ -far from $k$-junta:
  - main lemma $\rightarrow$ with probability $>5/6$ over choice of $\mathcal{G}$, $f$ is $\epsilon/2$-far from $k$-part junta w.r.t. $\mathcal{G}$
  - so, whp, if $\leq k$ parts found, $\inf([n]\setminus\text{found parts}) \geq \epsilon/2$
  - so $\Pr[\text{pick } x, y \text{ s.t. } f(x) \neq f(y_s x_s)] \geq \frac{\epsilon}{2}$
  - Expected tries to find such an $x,y$ pair $\leq \frac{2}{\epsilon}$
  - Expected tries to find $k+1$ relevant parts $\leq \frac{2}{\epsilon}(k+1)$
  - Markov’s Inequality $\rightarrow$ find $k+1$ relevant parts in $\frac{12}{\epsilon}(k+1)$ rounds with probability $5/6$
  - $\Pr[T \text{ passes}] < 1/6 + 1/6 = 1/3$
Proof of Lemma

- see board