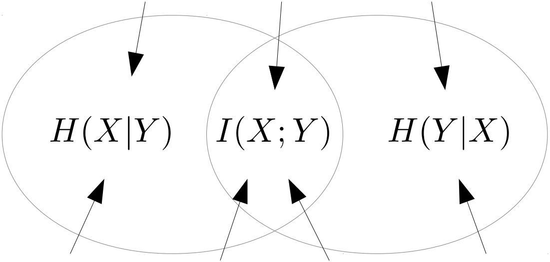
Review

$$H(X,Y) = H(X|Y) + I(X;Y) + H(Y|X)$$



$$H(X) = H(X|Y) + I(X;Y) \qquad I(X;Y) + H(Y|X) = H(Y)$$

Review

- Conditioning reduces entropy

Review

- Conditioning reduces entropy
- Chain rules:
  - $H(X, Y, Z)$

Review

- Conditioning reduces entropy
- Chain rules:
  - $H(X, Y, Z) = H(Z) + H(Y|X) + H(Z|X, Y)$
  - $H(X, Y, U|V)$

Review

- Conditioning reduces entropy
- Chain rules:
  - $H(X, Y, Z) = H(Z) + H(Y|X) + H(Z|X, Y)$
  - $H(X, Y, U|V) = H(X|V) + H(Y|X, V) + H(U|Y, X, V)$
  - $I(X, Y, Z; U)$

Review

- Conditioning reduces entropy
- Chain rules:
  - $H(X,Y,Z) = H(Z) \; + H(Y|X) \; + H(Z|X,Y)$
  - $H(X,Y,U|V) = H(X|V) + H(Y|X,V) + H(U|Y,X,V)$
  - $I(X,Y,Z;U) = I(X;U) + I(Y;U|X) + I(Z;U|X,Y)$
  - $I(X,Y,Z;U|V)$

Review

- Conditioning reduces entropy
- Chain rules:
  - $H(X,Y,Z) = H(Z) + H(Y|X) + H(Z|X,Y)$
  - $H(X,Y,U|V) = H(X|V) + H(Y|X,V) + H(U|Y,X,V)$
  - $I(X,Y,Z;U) = I(X;U) + I(Y;U|X) + I(Z;U|X,Y)$
  - $I(X,Y,Z;U|V) = I(X;U|V) + I(Y;U|V,X) + I(Z;U|V,X,Y)$
- Data processing inequality: if $X \perp Y|Z$,

Review

- Conditioning reduces entropy
- Chain rules:
  - $H(X,Y,Z) = H(Z) \; + H(Y|X) \; + H(Z|X,Y)$
  - $H(X,Y,U|V) = H(X|V) + H(Y|X,V) + H(U|Y,X,V)$
  - $I(X,Y,Z;U) = I(X;U) + I(Y;U|X) + I(Z;U|X,Y)$
  - $I(X,Y,Z;U|V) = I(X;U|V) + I(Y;U|V,X) + \; I(Z;U|V,X,Y)$
- Data processing inequality: if $X \perp Y|Z$, $I(X;Y) \geq I(X;Z)$
- Independence and mutual information:
  - $X \perp Y \Leftrightarrow$

Review

- Conditioning reduces entropy
- Chain rules:
  - $H(X,Y,Z) = H(Z) \ + H(Y|X) \ + H(Z|X,Y)$
  - $H(X,Y,U|V)= H(X|V) + H(Y|X,V) + H(U|Y,X,V)$
  - $I(X,Y,Z;U)= I(X;U) + I(Y;U|X) + I(Z;U|X,Y)$
  - $I(X,Y,Z;U|V)= I(X;U|V) + I(Y;U|V,X) + \ I(Z;U|V,X,Y)$
- Data processing inequality: if $X \perp Y|Z$, $I(X;Y) \geq I(X;Z)$
- Independence and mutual information:
  - $X \perp Y \Leftrightarrow I(X;Y) = 0$
  - $X \perp Y|Z \Leftrightarrow$

Review

- Conditioning reduces entropy
- Chain rules:
  - $H(X,Y,Z) = H(Z) + H(Y|X) + H(Z|X,Y)$
  - $H(X,Y,U|V) = H(X|V) + H(Y|X,V) + H(U|Y,X,V)$
  - $I(X,Y,Z;U) = I(X;U) + I(Y;U|X) + I(Z;U|X,Y)$
  - $I(X,Y,Z;U|V) = I(X;U|V) + I(Y;U|V,X) + I(Z;U|V,X,Y)$
- Data processing inequality: if $X \perp Y|Z$, $I(X;Y) \geq I(X;Z)$
- Independence and mutual information:
  - $X \perp Y \Leftrightarrow I(X;Y) = 0$
  - $X \perp Y|Z \Leftrightarrow I(X;Y|Z) = 0$
- KL-divergence: $KL(p||q) \triangleq$

Review

- Conditioning reduces entropy
- Chain rules:
  - $H(X,Y,Z) = H(Z) + H(Y|X) + H(Z|X,Y)$
  - $H(X,Y,U|V) = H(X|V) + H(Y|X,V) + H(U|Y,X,V)$
  - $I(X,Y,Z;U) = I(X;U) + I(Y;U|X) + I(Z;U|X,Y)$
  - $I(X,Y,Z;U|V) = I(X;U|V) + I(Y;U|V,X) + I(Z;U|V,X,Y)$
- Data processing inequality: if $X \perp Y|Z$, $I(X;Y) \geq I(X;Z)$
- Independence and mutual information:
  - $X \perp Y \Leftrightarrow I(X;Y) = 0$
  - $X \perp Y|Z \Leftrightarrow I(X;Y|Z) = 0$
- KL-divergence: $KL(p||q) \triangleq \sum_x p(x) \log \frac{p(x)}{q(x)}$

This time

- Identification/Decision trees
- Random forests
- Law of Large Number
- Asymptotic equipartition (AEP) and typical sequences

Vampire database

Romanian Data Base

Vampire?	Shadow?	Garlic?	Complexion?	Accent?
No	?	Yes	Pale	None
No	Yes	Yes	Ruddy	None
Yes	?	No	Ruddy	None
Yes	No	No	Average	Heavy
Yes	?	No	Average	Odd
No	Yes	No	Pale	Heavy
No	Yes	No	Average	Heavy
No	?	Yes	Ruddy	Odd

(https://www.youtube.com/watch?v=SXBG3RGr_Rc)

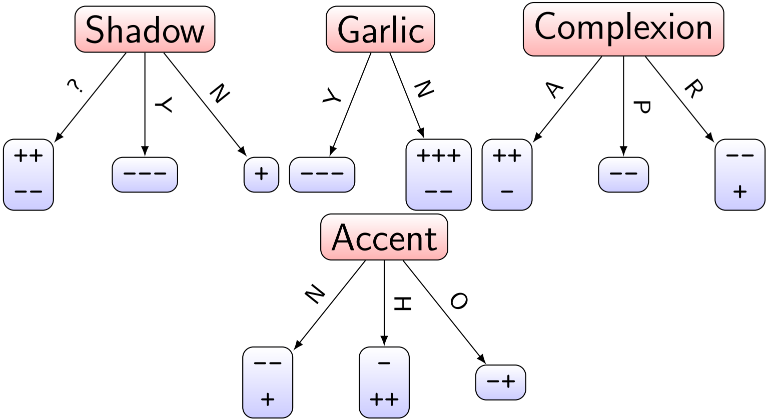
Identifying vampire

Goal: Design a set of tests to identify vampires

Potential difficulties

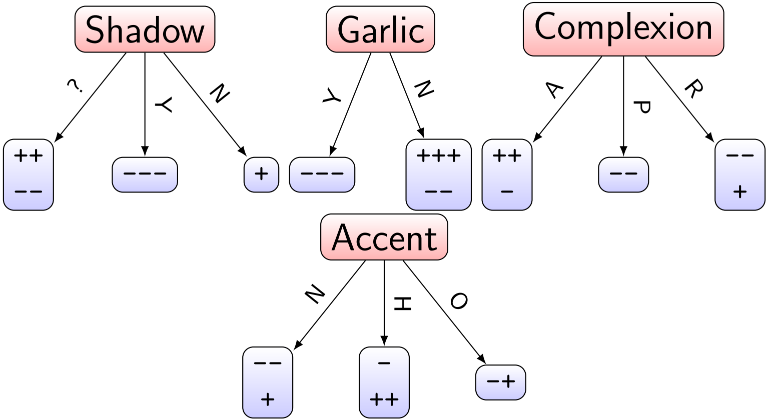
- Non-numerical data
- Some information may not matter
- Some may matter only sometimes
- Tests may be costly $\Rightarrow$ conduct as few as possible

Test trees



$+$: Vampire $\quad -$: Not vampire

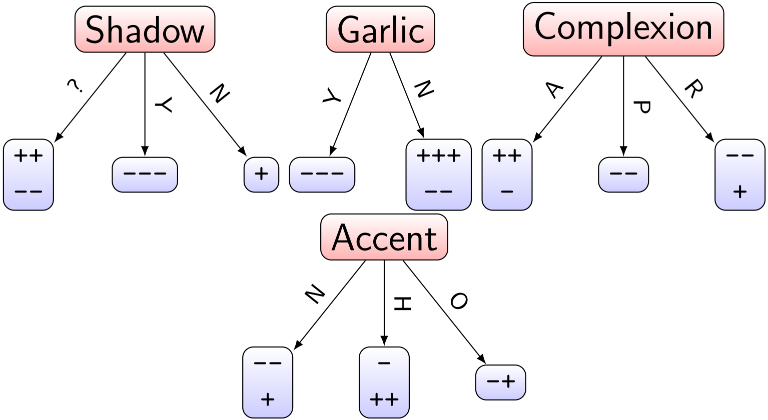
Test trees



$+$: Vampire $\quad -$: Not vampire

How to pick a good test?

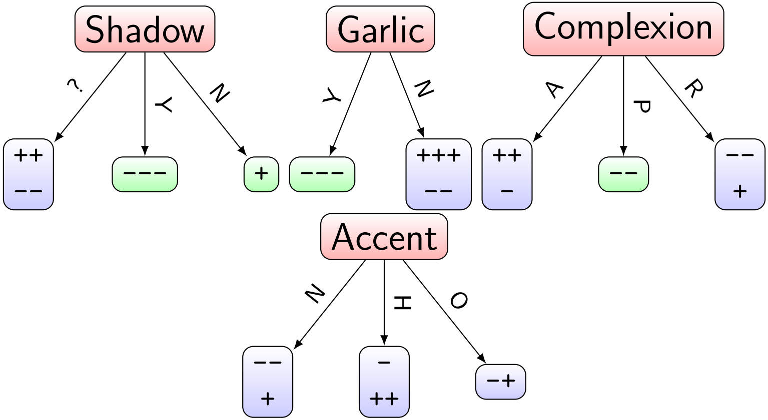
Test trees



$+$: Vampire $-$: Not vampire

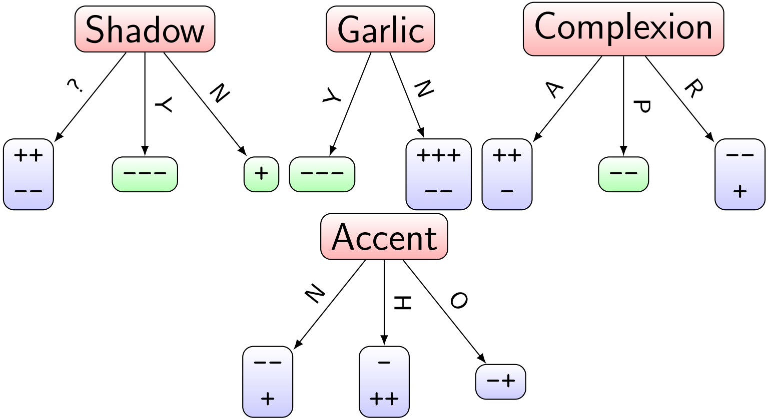
How to pick a good test? Pick test that identifies most vampires (and non-vampires)!

Sizes of homogeneous sets



$+$: Vampire $-$: Not vampire

Sizes of homogeneous sets

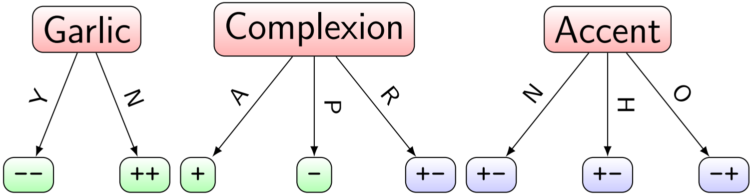


$+$: Vampire $-$: Not vampire

Shadow: 4 Garlic: 3 Complexion: 2 Accent: 0

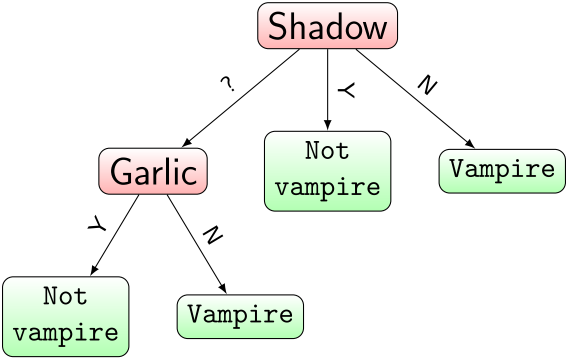
Picking second test

Let say we pick "shadow" as the first test after all. Then, for the remaining unclassified individuals,

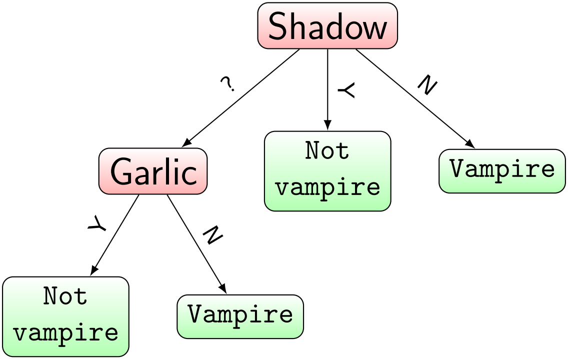


Garlic: 4 Complexion: 2 Accent: 0

Combined tests



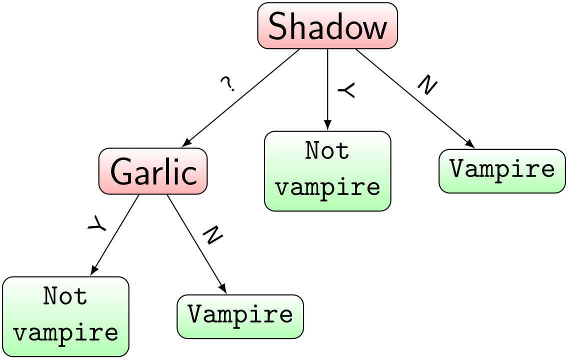
Combined tests



Problem

When our database size increases, none of the test likely to completely separate vampire from non-vampire. All tests will score 0 then.

Combined tests



Problem

When our database size increases, none of the test likely to completely separate vampire from non-vampire. All tests will score 0 then.

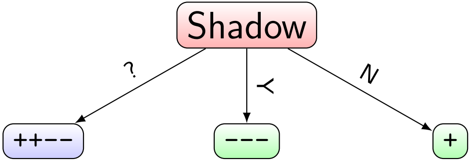
Entropy comes to the rescue!

Conditional entropy as a measure of test efficiency

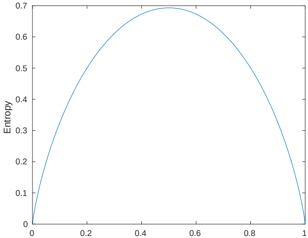
Consider the database is randomly sampled from a distribution. A set is

- Very homogeneous $\approx$ high certainty
- Not so homogenous $\approx$ high randomness

These can be measured with its entropy



$H(V|S=?)=1 \quad H(V|S=Y)=0 \quad H(V|S=N)=0$

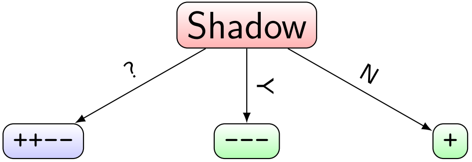


Conditional entropy as a measure of test efficiency

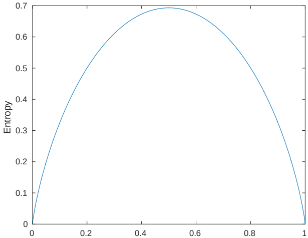
Consider the database is randomly sampled from a distribution. A set is

- Very homogeneous $\approx$ high certainty
- Not so homogenous $\approx$ high randomness

These can be measured with its entropy



$H(V|S=?)=1$ $H(V|S=Y)=0$ $H(V|S=N)=0$



Remaining uncertainty given the test:

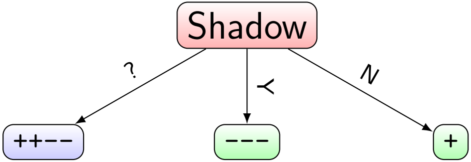
$$\frac{4}{8}H(V|S=?)$$

Conditional entropy as a measure of test efficiency

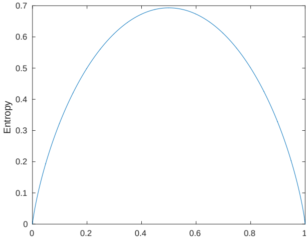
Consider the database is randomly sampled from a distribution. A set is

- Very homogeneous $\approx$ high certainty
- Not so homogenous $\approx$ high randomness

These can be measured with its entropy



$H(V|S=?)=1$ $H(V|S=Y)=0$ $H(V|S=N)=0$



Remaining uncertainty given the test:

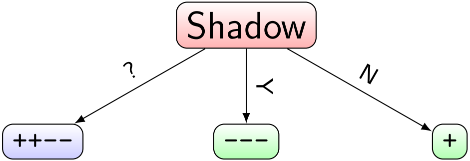
$$\frac{4}{8}H(V|S=?)+\frac{3}{8}H(V|S=Y)$$

Conditional entropy as a measure of test efficiency

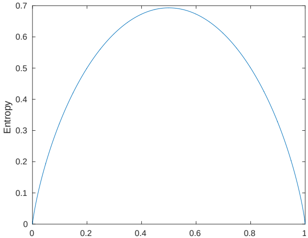
Consider the database is randomly sampled from a distribution. A set is

- Very homogeneous ≈ high certainty
- Not so homogenous ≈ high randomness

These can be measured with its entropy



$H(V|S=?)=1$ $\quad H(V|S=Y)=0$ $\quad H(V|S=N)=0$



Remaining uncertainty given the test:

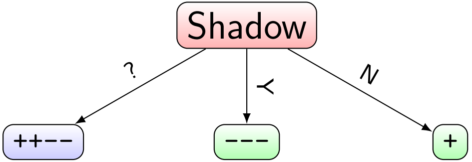
$$\frac{4}{8}H(V|S=?)+\frac{3}{8}H(V|S=Y)+\frac{1}{8}H(V|S=N)=0.5$$

Conditional entropy as a measure of test efficiency

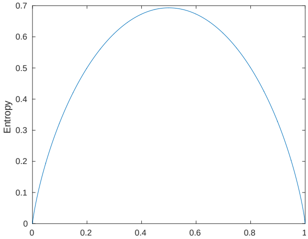
Consider the database is randomly sampled from a distribution. A set is

- Very homogeneous $\approx$ high certainty
- Not so homogenous $\approx$ high randomness

These can be measured with its entropy



$H(V|S=?)=1$ $H(V|S=Y)=0$ $H(V|S=N)=0$



Remaining uncertainty given the test:

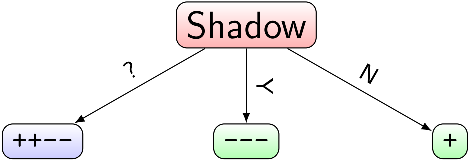
$$\frac{4}{8}H(V|S=?)+\frac{3}{8}H(V|S=Y)+\frac{1}{8}H(V|S=N)=0.5$$
$$=Pr(S=?)H(V|S=?)+Pr(S=Y)H(V|S=Y)+Pr(S=N)H(V|S=N)$$

Conditional entropy as a measure of test efficiency

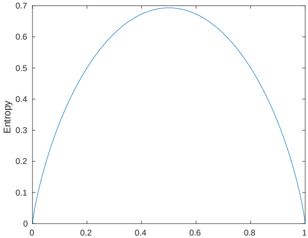
Consider the database is randomly sampled from a distribution. A set is

- Very homogeneous $\approx$ high certainty
- Not so homogenous $\approx$ high randomness

These can be measured with its entropy



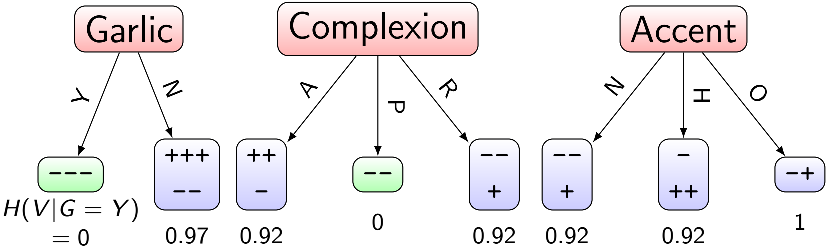
$H(V|S=?) = 1 \quad H(V|S=Y) = 0 \quad H(V|S=N) = 0$



Remaining uncertainty given the test:

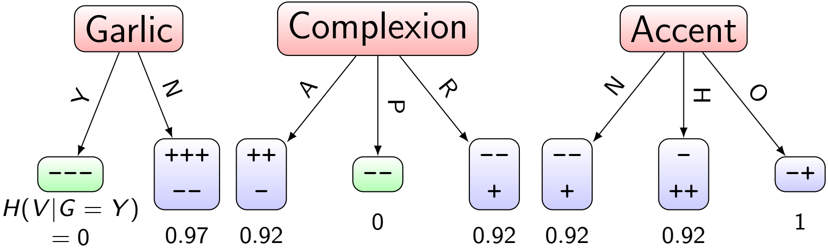
$$
\begin{aligned}
&\frac{4}{8}H(V|S=?) + \frac{3}{8}H(V|S=Y) + \frac{1}{8}H(V|S=N) = 0.5\\
=&Pr(S=?)H(V|S=?) + Pr(S=Y)H(V|S=Y) + Pr(S=N)H(V|S=N)\\
=&H(V|S)
\end{aligned}
$$

Remaining uncertainty



$H(V|S) = 0.5$

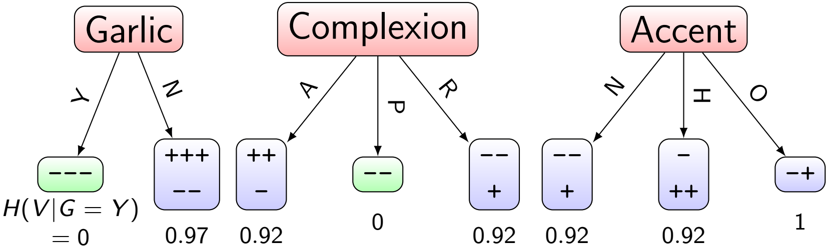
Remaining uncertainty



$$H(V|S) = 0.5$$

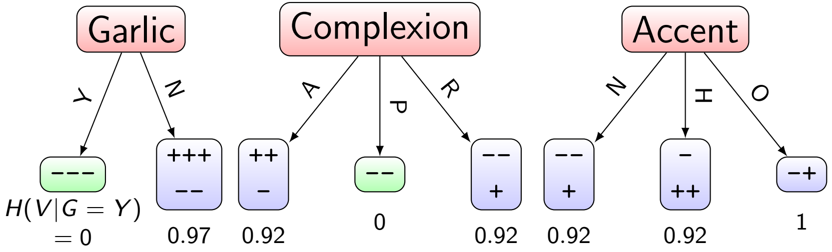
$$H(V|G) = \frac{3}{8}\cdot 0 + \frac{5}{8}\cdot 0.97 = 0.61$$

Remaining uncertainty



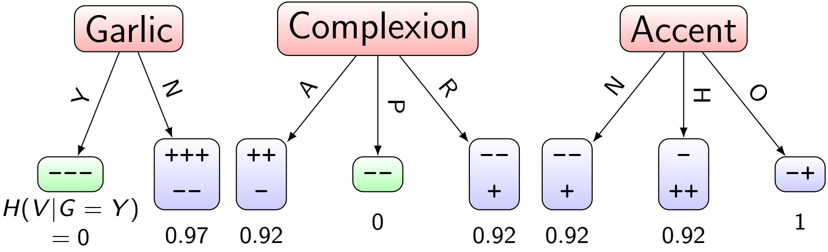
$$
\begin{aligned}
H(V|S) &= 0.5 \\
H(V|G) &= \frac{3}{8} \cdot 0 + \frac{5}{8} \cdot 0.97 = 0.61 \\
H(V|C) &= \frac{3}{8} \cdot 0.92 + \frac{2}{8} \cdot 0 + \frac{3}{8} \cdot 0.92 = 0.69
\end{aligned}
$$

Remaining uncertainty



$$\begin{aligned}
H(V|S) &= 0.5 \\
H(V|G) &= \frac{3}{8}\cdot 0 + \frac{5}{8}\cdot 0.97 = 0.61 \\
H(V|C) &= \frac{3}{8}\cdot 0.92 + \frac{2}{8}\cdot 0 + \frac{3}{8}\cdot 0.92 = 0.69 \\
H(V|A) &= \frac{3}{8}\cdot 0.92 + \frac{3}{8}\cdot 0.92 + \frac{2}{8}\cdot 1 = 0.94
\end{aligned}$$

Remaining uncertainty



$$H(V|S) = 0.5$$

$$H(V|G) = \frac{3}{8} \cdot 0 + \frac{5}{8} \cdot 0.97 = 0.61$$

$$H(V|C) = \frac{3}{8} \cdot 0.92 + \frac{2}{8} \cdot 0 + \frac{3}{8} \cdot 0.92 = 0.69$$

$$H(V|A) = \frac{3}{8} \cdot 0.92 + \frac{3}{8} \cdot 0.92 + \frac{2}{8} \cdot 1 = 0.94$$

Order of tests to pick: $S \succ G \succ C \succ A$

Potential extensions

- The test does not need to return discrete result. Let X be the test outcome. It can be continuous as well

Potential extensions

- The test does not need to return discrete result. Let X be the test outcome. It can be continuous as well
  - We should just pick i such that $H(V|X_i)$ to be as small as possible

Potential extensions

- The test does not need to return discrete result. Let X be the test outcome. It can be continuous as well
  - We should just pick i such that $H(V|X_i)$ to be as small as possible
  - It is equivalent of saying $I(V;X_i) = H(V) - H(V|X_i)$ is as large as possible. This is intuitive because we want to pick the information that is most relevant (sharing most information with) to V

Potential extensions

- The test does not need to return discrete result. Let X be the test outcome. It can be continuous as well
  - We should just pick i such that $H(V|X_i)$ to be as small as possible
  - It is equivalent of saying $I(V;X_i) = H(V) - H(V|X_i)$ is as large as possible. This is intuitive because we want to pick the information that is most relevant (sharing most information with) to V
- Build a number of trees instead of a single tree $\Rightarrow$ random forests

Random forests

- Pick random subset of training samples
- Train on each random subset but limited to a subset of features/attributes
- Given a test sample
  - Classify sample using each of the trees
  - Make final decision based on majority vote

Law of Large Number (LLN)

If we randomly sample $x_1, x_2, \cdots, x_N$ from an i.i.d. (identical and independently distributed) source, the average of $f(x_i)$ will approach the expected value as $N \to \infty$. That is,

$$\frac{1}{N}\sum_{i=1}^{N} f(x_i) = E[f(X)] \qquad \text{as } N \to \infty$$

Law of Large Number (LLN)

If we randomly sample $x_1, x_2, \cdots, x_N$ from an i.i.d. (identical and independently distributed) source, the average of $f(x_i)$ will approach the expected value as $N \to \infty$. That is,

$$\frac{1}{N}\sum_{i=1}^{N} f(x_i) = E[f(X)] \qquad \text{as } N \to \infty$$

Example

This is precisely how poll supposes to work! Pollster randomly draws sample from a portion of the population but will expect the prediction matches the outcome

Proof of LLN

The LLN is a rather strong result. We will only show a weak version here

$$Pr\left(\left|\frac{1}{N}\sum_{i=1}^{N} f(X_i) - E[f(X)]\right| \geq a\right) \leq \frac{Var(f(X))}{Na^2} \propto \frac{1}{N}$$

Proof of LLN

The LLN is a rather strong result. We will only show a weak version here

$$Pr\left(\left|\frac{1}{N}\sum_{i=1}^{N} f(X_i) - E[f(X)]\right| \geq a\right) \leq \frac{Var(f(X))}{Na^2} \propto \frac{1}{N}$$

Markov's Inequality

$$Pr(X \geq b) \leq \frac{E[X]}{b} \qquad \text{if } X \geq 0$$

Proof of LLN

The LLN is a rather strong result. We will only show a weak version here

$$Pr\left(\left|\frac{1}{N}\sum_{i=1}^{N} f(X_i) - E[f(X)]\right| \geq a\right) \leq \frac{Var(f(X))}{Na^2} \propto \frac{1}{N}$$

Markov's Inequality

$$Pr(X \geq b) \leq \frac{E[X]}{b} \qquad \text{if } X \geq 0$$

Proof:

$X = I(X \geq b) \cdot X + I(X < b) \cdot X \geq I(X \geq b) \cdot b$

Proof of LLN

The LLN is a rather strong result. We will only show a weak version here

$$Pr\left(\left|\frac{1}{N}\sum_{i=1}^{N} f(X_i) - E[f(X)]\right| \geq a\right) \leq \frac{Var(f(X))}{Na^2} \propto \frac{1}{N}$$

Markov's Inequality

$$Pr(X \geq b) \leq \frac{E[X]}{b} \qquad \text{if } X \geq 0$$

Proof:

$X = I(X \geq b) \cdot X + I(X < b) \cdot X \geq I(X \geq b) \cdot b \Rightarrow \quad E[X] \geq Pr(X \geq b) \cdot b$

Proof of LLN

Markov's Inequality

$$Pr(X \geq b) \leq \frac{E[X]}{b} \qquad \text{if } X \geq 0$$

Chebyshev's Inequality

$$Pr(|Y - E[Y]| \geq a) \leq \frac{Var(Y)}{a^2}$$

Proof of LLN

Markov's Inequality

$$Pr(X \geq b) \leq \frac{E[X]}{b} \qquad \text{if } X \geq 0$$

Chebyshev's Inequality

$$Pr(|Y - E[Y]| \geq a) \leq \frac{Var(Y)}{a^2}$$

Proof: Take $X = |Y - E[Y]|^2$ and $b = a^2$, by Markov's Inequality

Proof of LLN

Markov's Inequality

$$Pr(X \geq b) \leq \frac{E[X]}{b} \qquad \text{if } X \geq 0$$

Chebyshev's Inequality

$$Pr(|Y - E[Y]| \geq a) \leq \frac{Var(Y)}{a^2}$$

Proof: Take $X = |Y - E[Y]|^2$ and $b = a^2$, by Markov's Inequality

$$Pr(|Y - E[Y]| \geq a) = Pr(|Y - E[Y]|^2 \geq a^2)$$

$$\leq \frac{E[|Y - E[Y]|^2]}{a^2}$$

Proof of LLN

Markov's Inequality

$$Pr(X \geq b) \leq \frac{E[X]}{b} \qquad \text{if } X \geq 0$$

Chebyshev's Inequality

$$Pr(|Y - E[Y]| \geq a) \leq \frac{Var(Y)}{a^2}$$

Proof: Take $X = |Y - E[Y]|^2$ and $b = a^2$, by Markov's Inequality

$$Pr(|Y - E[Y]| \geq a) = Pr(|Y - E[Y]|^2 \geq a^2)$$

$$\leq \frac{E[|Y - E[Y]|^2]}{a^2} = \frac{Var(Y)}{a^2}$$

Proof of LLN

Chebyshev's Inequality

$$Pr(|Y - E[Y]| \geq a) \leq \frac{Var(Y)}{a^2}$$

Proof of weak LLN

Let $Z = \frac{1}{N}\sum_{i=1}^{N} f(X_i)$, apparently $E[Z] = E[f(X)]$ and

$$Var(Z) = \frac{1}{N^2}\sum_{i=1}^{N} Var(f(X)) = \frac{Var(f(X))}{N}$$

By Chebyshev's Inequality,

Proof of LLN

Chebyshev's Inequality

$$Pr(|Y - E[Y]| \geq a) \leq \frac{Var(Y)}{a^2}$$

Proof of weak LLN

Let $Z = \frac{1}{N}\sum_{i=1}^{N} f(X_i)$, apparently $E[Z] = E[f(X)]$ and

$$Var(Z) = \frac{1}{N^2}\sum_{i=1}^{N} Var(f(X)) = \frac{Var(f(X))}{N}$$

By Chebyshev's Inequality,

$$Pr\left(\left|\frac{1}{N}\sum_{i=1}^{N} f(X_i) - E[f(X)]\right| \geq a\right)$$

$$= Pr(|Z - E[Z]| \geq a) \leq \frac{Var(Z)}{a^2}$$

Proof of LLN

Chebyshev's Inequality

$$Pr(|Y - E[Y]| \geq a) \leq \frac{Var(Y)}{a^2}$$

Proof of weak LLN

Let $Z = \frac{1}{N}\sum_{i=1}^{N} f(X_i)$, apparently $E[Z] = E[f(X)]$ and

$$Var(Z) = \frac{1}{N^2}\sum_{i=1}^{N} Var(f(X)) = \frac{Var(f(X))}{N}$$

By Chebyshev's Inequality,

$$Pr\left(\left|\frac{1}{N}\sum_{i=1}^{N} f(X_i) - E[f(X)]\right| \geq a\right)$$

$$= Pr(|Z - E[Z]| \geq a) \leq \frac{Var(Z)}{a^2} = \frac{Var(f(X))}{Na^2}$$

Main idea

Consider a sequence of symbols $x_1, x_2, \cdots, x_N$ sampled from a DMS and consider the sample average of the log-probabilities of each sampled symbols

$$\frac{1}{N}\sum_{i=1}^{N}\log\frac{1}{p(x_i)} \to E\left[\log\frac{1}{p(X)}\right]$$

by LLN.

Main idea

Consider a sequence of symbols $x_1, x_2, \cdots, x_N$ sampled from a DMS and consider the sample average of the log-probabilities of each sampled symbols

$$\frac{1}{N}\sum_{i=1}^{N} \log \frac{1}{p(x_i)} \to E\left[\log \frac{1}{p(X)}\right] = H(X)$$

by LLN.

Main idea

Consider a sequence of symbols $x_1, x_2, \cdots, x_N$ sampled from a DMS and consider the sample average of the log-probabilities of each sampled symbols

$$\frac{1}{N}\sum_{i=1}^{N} \log \frac{1}{p(x_i)} \to E\left[\log \frac{1}{p(X)}\right] = H(X)$$

by LLN. But for the LHS,

$$\frac{1}{N}\sum_{i=1}^{N} \log \frac{1}{p(x_i)} = \frac{1}{N} \log \frac{1}{\prod_{i=1}^{N} p(x_i)} = -\frac{1}{N} \log p(x^N),$$

where $x^N = x_1, x_2, \cdots, x_N$

Main idea

Consider a sequence of symbols $x_1, x_2, \cdots, x_N$ sampled from a DMS and consider the sample average of the log-probabilities of each sampled symbols

$$\frac{1}{N}\sum_{i=1}^{N} \log \frac{1}{p(x_i)} \to E\left[\log \frac{1}{p(X)}\right] = H(X)$$

by LLN. But for the LHS,

$$\frac{1}{N}\sum_{i=1}^{N} \log \frac{1}{p(x_i)} = \frac{1}{N} \log \frac{1}{\prod_{i=1}^{N} p(x_i)} = -\frac{1}{N} \log p(x^N),$$

where $x^N = x_1, x_2, \cdots, x_N$

Rearranging the terms, this implies that for any sequence sampled from the source, the probability of the sampled sequence $p(x^N) \to 2^{-NH(X)}$!

Set of typical sequences

Let's name the sequence x^N with $p(x^N) \sim 2^{-NH(X)}$ typical and define the set of typical sequences

$$\mathcal{A}_\epsilon^N(X) = \{x^N | 2^{-N(H(X)+\epsilon)} \leq p(x^N) \leq 2^{-N(H(X)-\epsilon)}\}$$

Set of typical sequences

Let's name the sequence x^N with $p(x^N) \sim 2^{-NH(X)}$ typical and define the set of typical sequences

$$\mathcal{A}_\epsilon^N(X) = \{x^N | 2^{-N(H(X)+\epsilon)} \leq p(x^N) \leq 2^{-N(H(X)-\epsilon)}\}$$

- For any $\epsilon > 0$, we can find a sufficiently large N such that any sampled sequence from the source is typical

Set of typical sequences

Let's name the sequence x^N with $p(x^N) \sim 2^{-NH(X)}$ typical and define the set of typical sequences

$$\mathcal{A}_\epsilon^N(X) = \{x^N | 2^{-N(H(X)+\epsilon)} \leq p(x^N) \leq 2^{-N(H(X)-\epsilon)}\}$$

- For any $\epsilon > 0$, we can find a sufficiently large N such that any sampled sequence from the source is typical
- Since all typical sequences have probability $\sim 2^{-NH(X)}$ and they fill up the entire probability space (everything is typical), there should be approximately $\frac{1}{2^{-NH(X)}} = 2^{NH(X)}$ typical sequences

Precise bounds on the size of typical set

$$(1-\delta)2^{N(H(X)-\epsilon)} \leq |\mathcal{A}_\epsilon^N(X)| \leq 2^{N(H(X)+\epsilon)}$$

$$1 \geq Pr(X^N \in \mathcal{A}_\epsilon^N(X))$$

Precise bounds on the size of typical set

$$(1-\delta)2^{N(H(X)-\epsilon)} \leq |\mathcal{A}_\epsilon^N(X)| \leq 2^{N(H(X)+\epsilon)}$$

$$1 \geq Pr(X^N \in \mathcal{A}_\epsilon^N(X)) = \sum_{x^N \in \mathcal{A}_\epsilon^N(X)} p(x^N)$$

Precise bounds on the size of typical set

$$(1-\delta)2^{N(H(X)-\epsilon)} \leq |\mathcal{A}_\epsilon^N(X)| \leq 2^{N(H(X)+\epsilon)}$$

$$1 \geq Pr(X^N \in \mathcal{A}_\epsilon^N(X)) = \sum_{x^N \in \mathcal{A}_\epsilon^N(X)} p(x^N) \geq \sum_{x^N \in \mathcal{A}_\epsilon^N(X)} 2^{-N(H(X)+\epsilon)}$$

Precise bounds on the size of typical set

$$(1-\delta)2^{N(H(X)-\epsilon)} \leq |\mathcal{A}_\epsilon^N(X)| \leq 2^{N(H(X)+\epsilon)}$$

$$1 \geq Pr(X^N \in \mathcal{A}_\epsilon^N(X)) = \sum_{x^N \in \mathcal{A}_\epsilon^N(X)} p(x^N) \geq \sum_{x^N \in \mathcal{A}_\epsilon^N(X)} 2^{-N(H(X)+\epsilon)}$$
$$= |\mathcal{A}_\epsilon^N(X)| 2^{-N(H(X)+\epsilon)}$$

Precise bounds on the size of typical set

$$(1-\delta)2^{N(H(X)-\epsilon)} \le |\mathcal{A}_\epsilon^N(X)| \le 2^{N(H(X)+\epsilon)}$$

$$1 \ge Pr(X^N \in \mathcal{A}_\epsilon^N(X)) = \sum_{x^N \in \mathcal{A}_\epsilon^N(X)} p(x^N) \ge \sum_{x^N \in \mathcal{A}_\epsilon^N(X)} 2^{-N(H(X)+\epsilon)}$$
$$= |\mathcal{A}_\epsilon^N(X)| 2^{-N(H(X)+\epsilon)}$$

For a sufficiently large N, we have

$$1-\delta \le Pr(X^N \in \mathcal{A}_\epsilon^N(X))$$

Precise bounds on the size of typical set

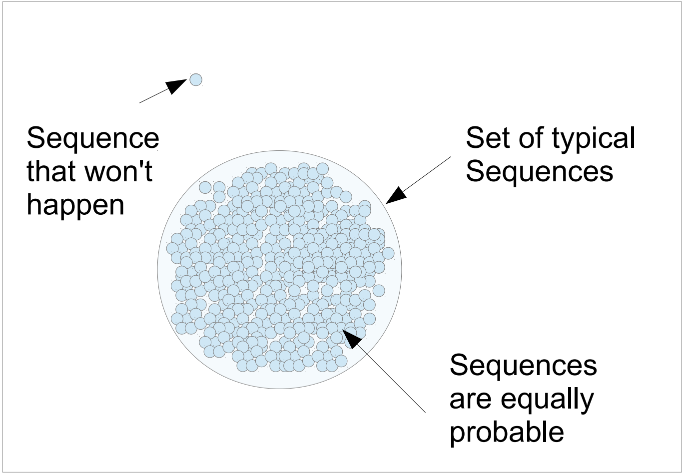
$$(1-\delta)2^{N(H(X)-\epsilon)} \leq |\mathcal{A}_\epsilon^N(X)| \leq 2^{N(H(X)+\epsilon)}$$

$$1 \geq Pr(X^N \in \mathcal{A}_\epsilon^N(X)) = \sum_{x^N \in \mathcal{A}_\epsilon^N(X)} p(x^N) \geq \sum_{x^N \in \mathcal{A}_\epsilon^N(X)} 2^{-N(H(X)+\epsilon)}$$
$$= |\mathcal{A}_\epsilon^N(X)| 2^{-N(H(X)+\epsilon)}$$

For a sufficiently large N, we have

$$1-\delta \leq Pr(X^N \in \mathcal{A}_\epsilon^N(X)) = \sum_{x^N \in \mathcal{A}_\epsilon^N(X)} p(x^N) \leq \sum_{x^N \in \mathcal{A}_\epsilon^N(X)} 2^{-N(H(X)-\epsilon)}$$
$$= |\mathcal{A}_\epsilon^N(X)| 2^{-N(H(X)-\epsilon)}$$

AEP



Asymptotic equipatition refers to the fact that the probability space is equally partitioned by the typical sequences

AEP and compression limit

Consider coin flipping again, let say $Pr(Head) = 0.3$ and $N = 1000$

AEP and compression limit

Consider coin flipping again, let say $Pr(Head) = 0.3$ and $N = 1000$

- The typical sequences will be those with approximately 300 heads and 700 tails

AEP and compression limit

Consider coin flipping again, let say $Pr(Head) = 0.3$ and $N = 1000$

- The typical sequences will be those with approximately 300 heads and 700 tails
- AEP (LLN) tells us that it is almost impossible to get, say, a sequence of 100 heads and 900 tails

AEP and compression limit

Consider coin flipping again, let say $Pr(Head) = 0.3$ and $N = 1000$

- The typical sequences will be those with approximately 300 heads and 700 tails
- AEP (LLN) tells us that it is almost impossible to get, say, a sequence of 100 heads and 900 tails
- AEP also tells us that the number of typical sequences are approximately $2^{NH(X)}$

AEP and compression limit

Consider coin flipping again, let say $Pr(Head) = 0.3$ and $N = 1000$

- The typical sequences will be those with approximately 300 heads and 700 tails
- AEP (LLN) tells us that it is almost impossible to get, say, a sequence of 100 heads and 900 tails
- AEP also tells us that the number of typical sequences are approximately $2^{NH(X)}$
- Therefore, we can simply assign index to all the typical sequences and ignore the rest. Then we only need $\log 2^{NH(X)} = NH(X)$ to store a sequence of N symbols. And on average, we need $H(X)$ bits per symbol as before!