Foundations of Trustworthy AI

Jonathan Ullman

CS7180/7880 · Fall 2026

Class 15: Finish memorization; Start membership inference

Oct 30, 2026

Information Theory 101

Name	Notation	Coin Toss Analogy
Entropy of U	$H(U)$	
Conditional entropy of U given V	$H(U\|V)$	
Mutual information between U and V	$I(U;V) = H(U) - H(U\|V)$ $I(U;V) = I(V;U)$	
Conditional mutual information btw U and V given W	$I(U;V\|W)$ $= H(U\|W) - H(U\|V,W)$	

Memorization of Irrelevant Data

= random variable

= known distribution

Q — prior (distribution over problems)

↓

D — distribution

↓

$S = \{(X_1, Y_1), \ldots, (X_m, Y_m)\}$ — training data

↓

$A(S) = \hat{\theta}$ — model (good for D)

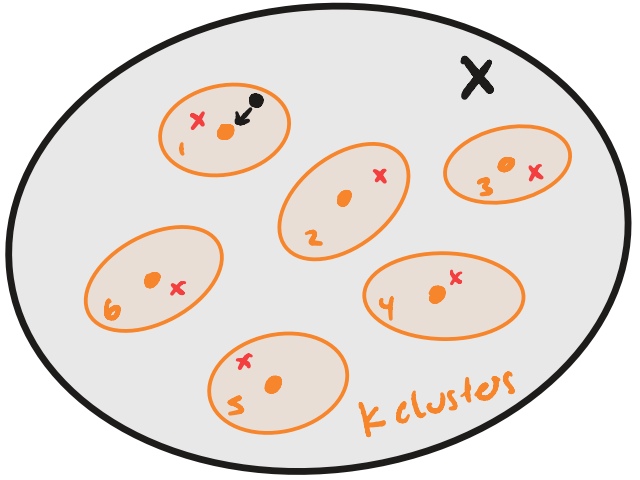
Goal: Construct a prior Q s.t. any "nearly optimal" model memorizes a lot of S

~~$I(S; \hat{\theta}) \approx H(S)$~~

$I(S; \hat{\theta} | D) \approx H(S|D)$

How much does the model memorize about S that is "irrelevant" to the distribution

A Hard Learning Problem



$X = \{0,1\}^d$ $\quad Y = \{1, \dots, n\}$

optimal classifier is nearest neighbor among the cluster centers

given one sample from each cluster nearest neighbor still does well

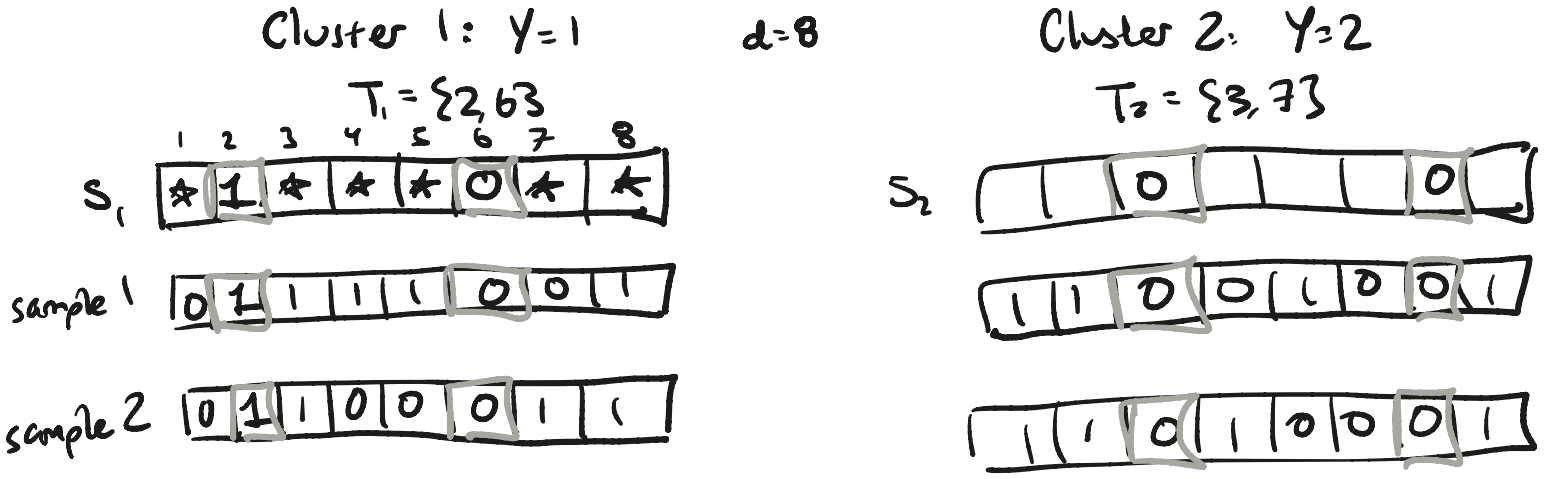
Defining Q / Choosing D:

for each cluster $j = 1 \dots n$:

[choose a set $T_j \subseteq [d]$, $s_j \in \{0,1\}^d$ → s_j is uniformly random; T_j contains coordinate l w.p. p

choose $Y = j$ at random, choose $X_i = \begin{cases} s_{j,i} & \text{if } i \in T_j \\ b_i, b_i \sim Be(\frac{1}{2}) & \text{if } i \notin T_j \end{cases}$

Cluster Classification Intuition



any one sample from a cluster is uniformly random in $\{0,1\}^d$

Cluster Classification

Given 1 sample from cluster j: (X, j)

A new point $\tilde{X}$ from some cluster j':

if $j' \neq j$ then $\tilde{X}$ is random and $\mathbb{E}[agr(X, \tilde{X})] = \frac{d}{2}$

if $j' = j$ then $\tilde{X}$ agrees on T_j and $\mathbb{E}[agr(X, \tilde{X})] = \frac{(1+\rho)d}{2}$

set ρ and d so that we can tell these apart and $\rho \to 0$ $d \to \infty$

WTS: Any accurate classifier has to memorize almost all of X even though only a tiny part of X is relevant to D.