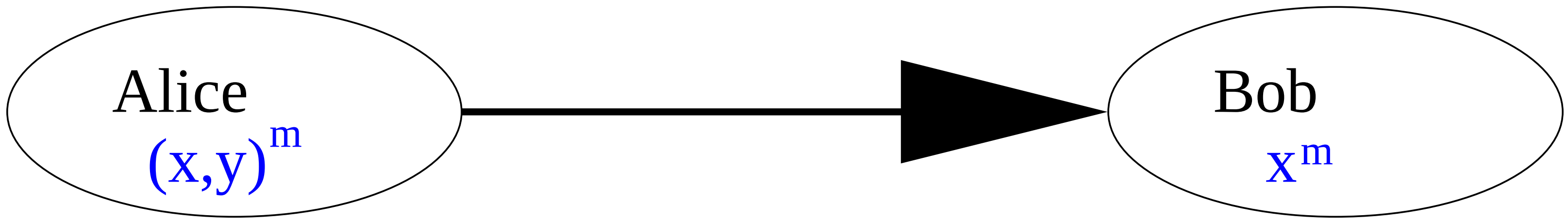
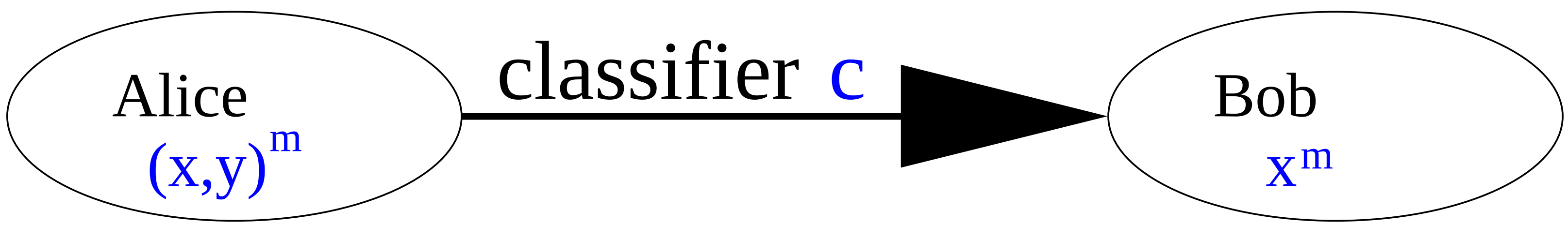
General Generalization Bounds, Part I

John Langford

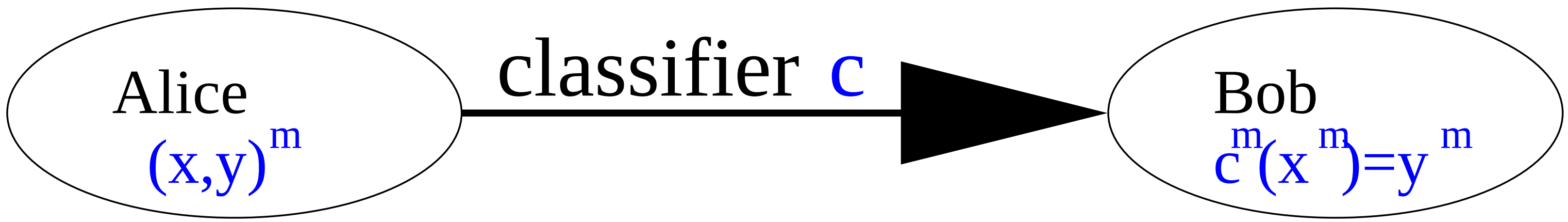
Theorem: (Occam's Razor Bound) Small description length for classifier $\Rightarrow$ verifiable learning



Theorem: (Occam's Razor Bound) Small description length for classifier $\Rightarrow$ verifiable learning

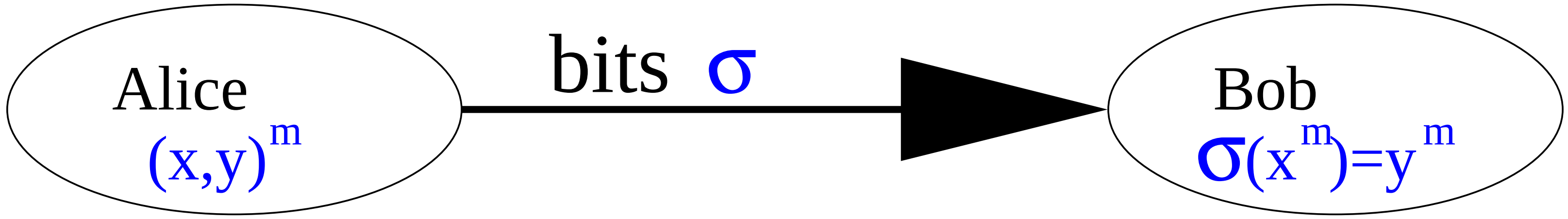


Theorem: (Occam's Razor Bound) Small description length for classifier $\Rightarrow$ verifiable learning



Restated: Small number of bits transmitted $\Rightarrow$verifiable learning

Suppose we transmitted the labels using some other language. Does this still hold?



Alice and Bob sort the unlabeled data

Alice sends "linear seperator number 2047"

Bob labels his data according to #2047

Does this method overfit?

Note: σ is NOT a classifier (what does "overfit" mean?)

Outline

1. For *transduction* label compression $\Rightarrow$ verifiable learning

2. Bound application to clustering

Basic Definitions

X = input space

$Y = \{0, 1\}$ = output space

$D(x, y)$ = distribution over X and Y

$S = (x, y)^m \sim D^m$ = train set

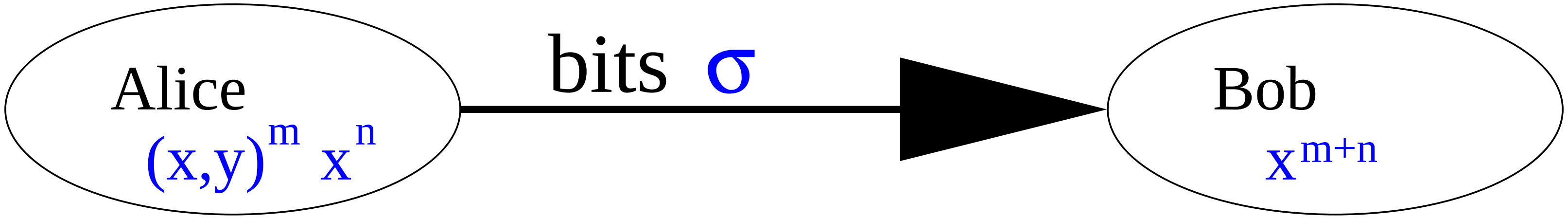
$S' = (x, y)^n \sim D^n$ = test set

$\sigma : X^{|S|+|S'|} \to Y^{|S|+|S'|}$ = transductive classifier

$\hat{\sigma}_S \equiv \sum_{x,y \in S} I(\sigma(S, S')(x) \neq y)$ = train error

$\hat{\sigma}_{S'} \equiv \sum_{x,y \in S'} I(\sigma(S, S')(x) \neq y)$ = test error

The compression game



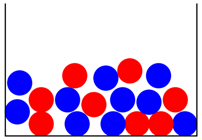
How many bits are used to send the labels to Bob?

PAC-MDL Bound: Simplified

Theorem: (PAC-MDL Bound, realizable case) For all description languages over σ, with probability $1-\delta$ over draws of $S, S' \sim D^{m+n}$:

$$\forall \sigma : \; \hat{\sigma}_S > 0 \text{ or } \hat{\sigma}_{S'} \leq \frac{|\sigma| + \log_2 \frac{1}{\delta}}{\log_2 \left(1 + \frac{m}{n}\right)}$$

Exact Bound

Given: a bucket  with 8 red balls and 10 blue balls

Pick: 10 balls (uniformly at random without replacement)

Question: What is the probability of choosing 4 red balls and 6 blue balls?

Answer:

$$\frac{\binom{10}{6}\binom{8}{4}}{\binom{18}{10}}$$

Basic Definition

Bucket(m, n, a, b) = Probability of getting at least b blue balls in $a + b$ draws from a bucket with m red balls and n blue balls

$$= \sum_{t=b}^{a+b} \frac{\binom{m}{a+b-t}\binom{n}{t}}{\binom{m+n}{a+b}}$$

$$b_{\text{max}}(m, n, a, \delta) = \max\{b : \text{Bucket}(m, n, a, b) \geq \delta\}$$

Theorem: (PAC-MDL Bound) For all measures $P(\sigma)$, with probability $1-\delta$ over draws of $S, S' \sim D^{m+n}$:

$$\forall \sigma : \ \hat{\sigma}_{S'} \leq b_{\mathsf{max}}(m, n, \hat{\sigma}_S, P(\sigma)\delta)$$

Proof

Fix $(x, y)^{m+n} \Rightarrow \hat{\sigma}_{S,S'}$ total errors.

Let $\pi(m, n)$ be a random partition of $(x, y)^{m+n}$ into S and S'.

(Sampling without replacement = partioning)

$\Rightarrow \mathrm{Pr}_{S,S' \sim \pi(m,n)}(t \text{ test errors and } \hat{\sigma}_{S,S'} - t \text{ train errors}) = \frac{\binom{m}{\hat{\sigma}_{S,S'}-t}\binom{n}{t}}{\binom{m+n}{\hat{\sigma}_{S,S'}}}$

(Definition of Bucket)

$\Rightarrow \mathrm{Pr}_{S,S' \sim \pi(m,n)}(\mathrm{Bucket}(m, n, \hat{\sigma}_S, \hat{\sigma}_{S'}) < P(\sigma)\delta) < P(\sigma)\delta$

Proof, continued

(Union bound $P(A \text{ or } B) \leq P(A) + P(B)$)

$\Rightarrow \text{Pr}_{S,S' \sim \pi(m,n)}(\exists \sigma : \text{ Bucket}(m, n, \hat{\sigma}_S, \hat{\sigma}_{S'}) < P(\sigma)\delta) < \delta$

(Negation)

$\Rightarrow \text{Pr}_{S,S' \sim \pi(m,n)}(\forall \sigma : \text{ Bucket}(m, n, \hat{\sigma}_S, \hat{\sigma}_{S'}) \geq P(\sigma)\delta) \geq 1 - \delta$

(Taking worst case over $\hat{\sigma}_{S'}$)

$\Rightarrow \text{Pr}_{S,S' \sim \pi(m,n)}(\forall \sigma : \hat{\sigma}_{S'} \leq b\text{max}(m, n, \hat{\sigma}_S, P(\sigma)\delta)) \geq 1 - \delta$

(Correct for each $(x, y)^{m+n}$ implies truth in expectation)

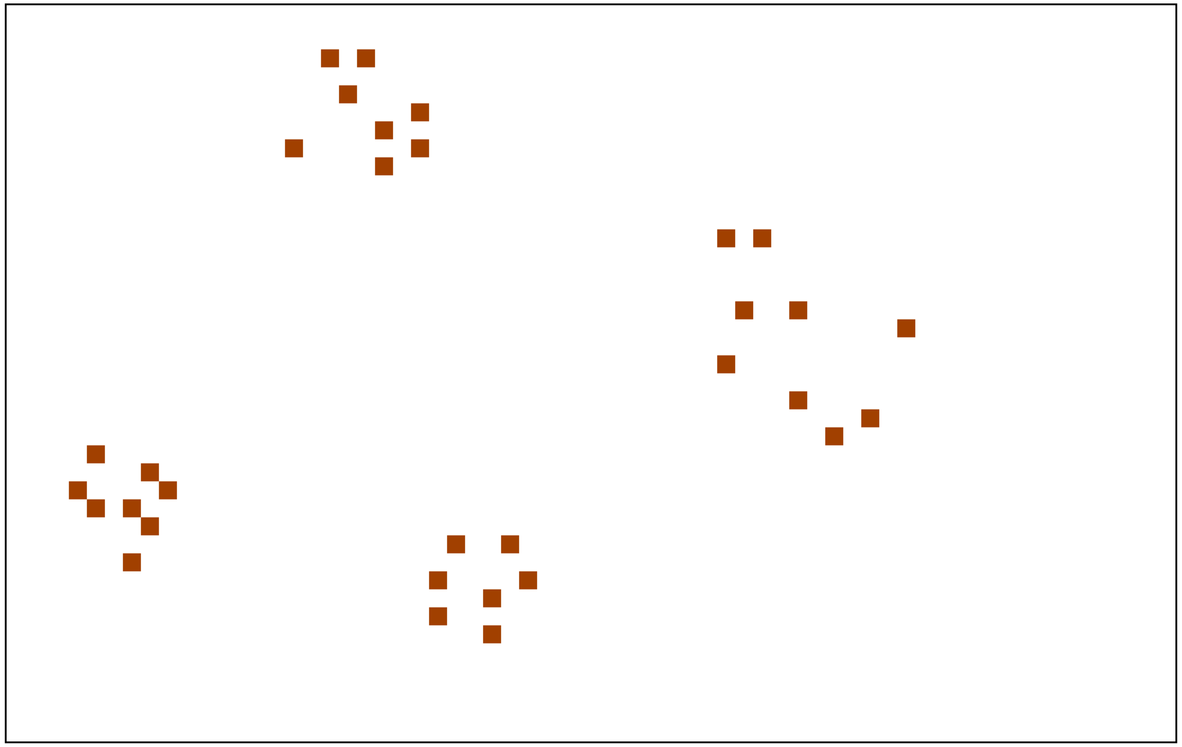
$\Rightarrow \text{Pr}_{S,S' \sim D^{m+n}}(\forall \sigma : \hat{\sigma}_{S'} \leq b\text{max}(m, n, \hat{\sigma}_S, P(\sigma)\delta)) \geq 1 - \delta$

Outline

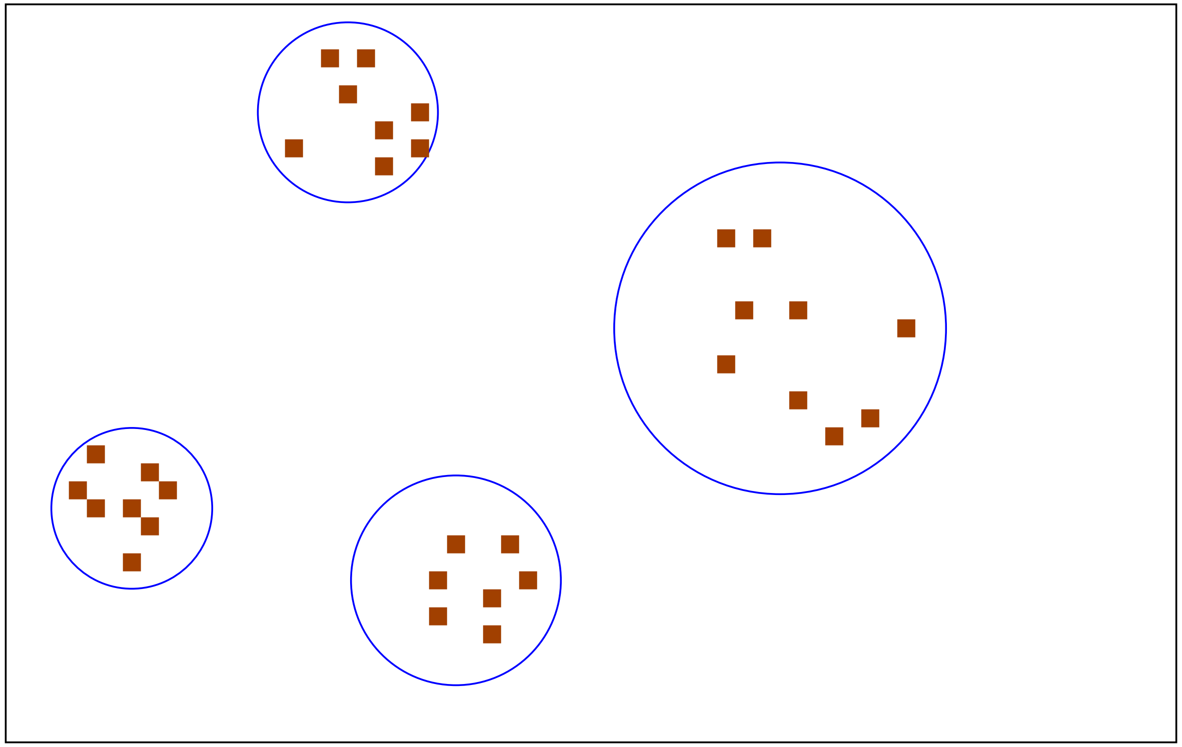
1. For *transduction* label compression $\Rightarrow$ verifiable learning

2. Bound application to clustering

A Set of Points



A Set of Clustered Points



Clustering

Definition: Given datapoints z^t, find a partition into k sets optimizing some criteria.

Common criteria:

1. Minimize {average, maximal} intracluster distance.

2. Maximize probability under some probabilistic model.

3. Many other criteria...

4. PAC-MDL bound (how useful is clustering for prediction?)

Clustering application

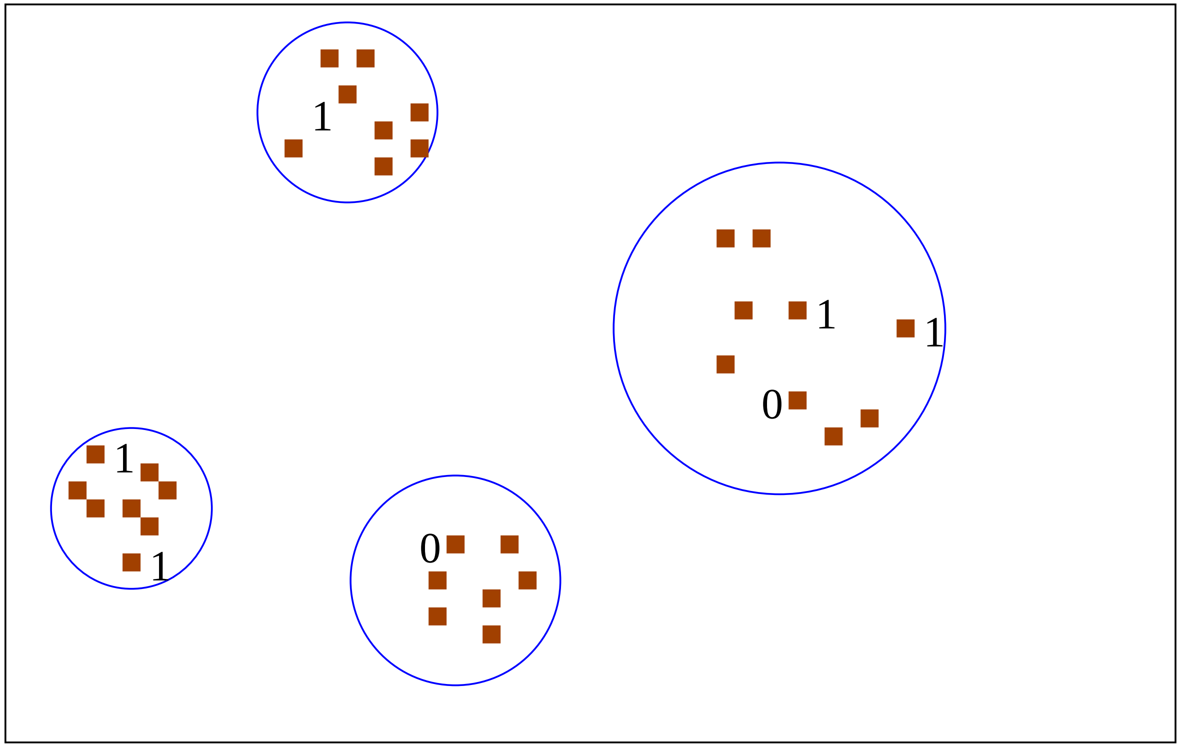
Suppose $z^t = (x, y)^m, x^n$

1. Cluster on x^{m+n}

2. Choose best constant label y for each cluster using y^m.

3. For each x, predict the label of it's cluster.

A Set of Clustered Points, With Labels



What is $P(\sigma)$?

What is $P(\sigma)$?

k = number of clusters

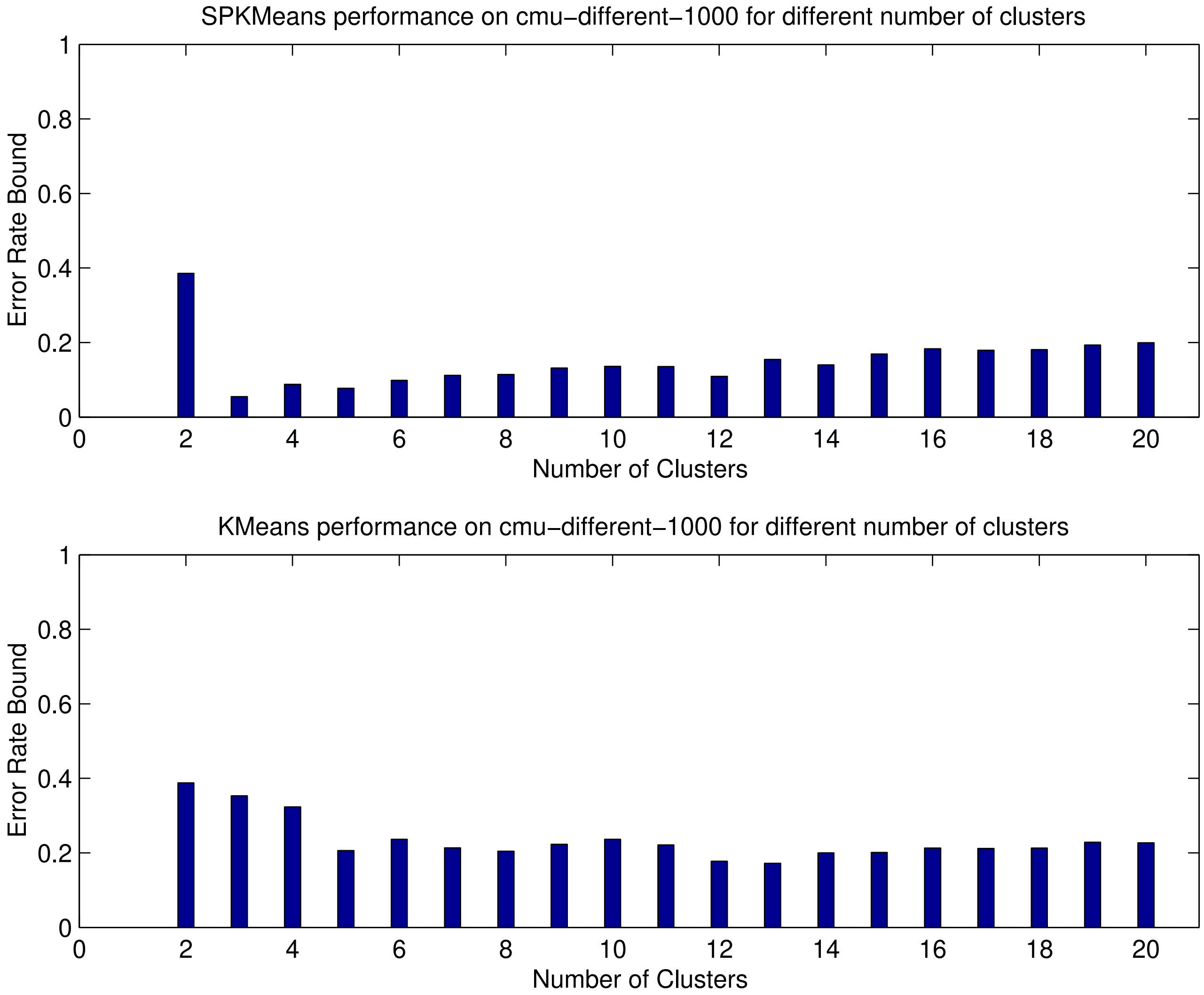
2 = number of labels

$P(\sigma) = 2^{-k}$

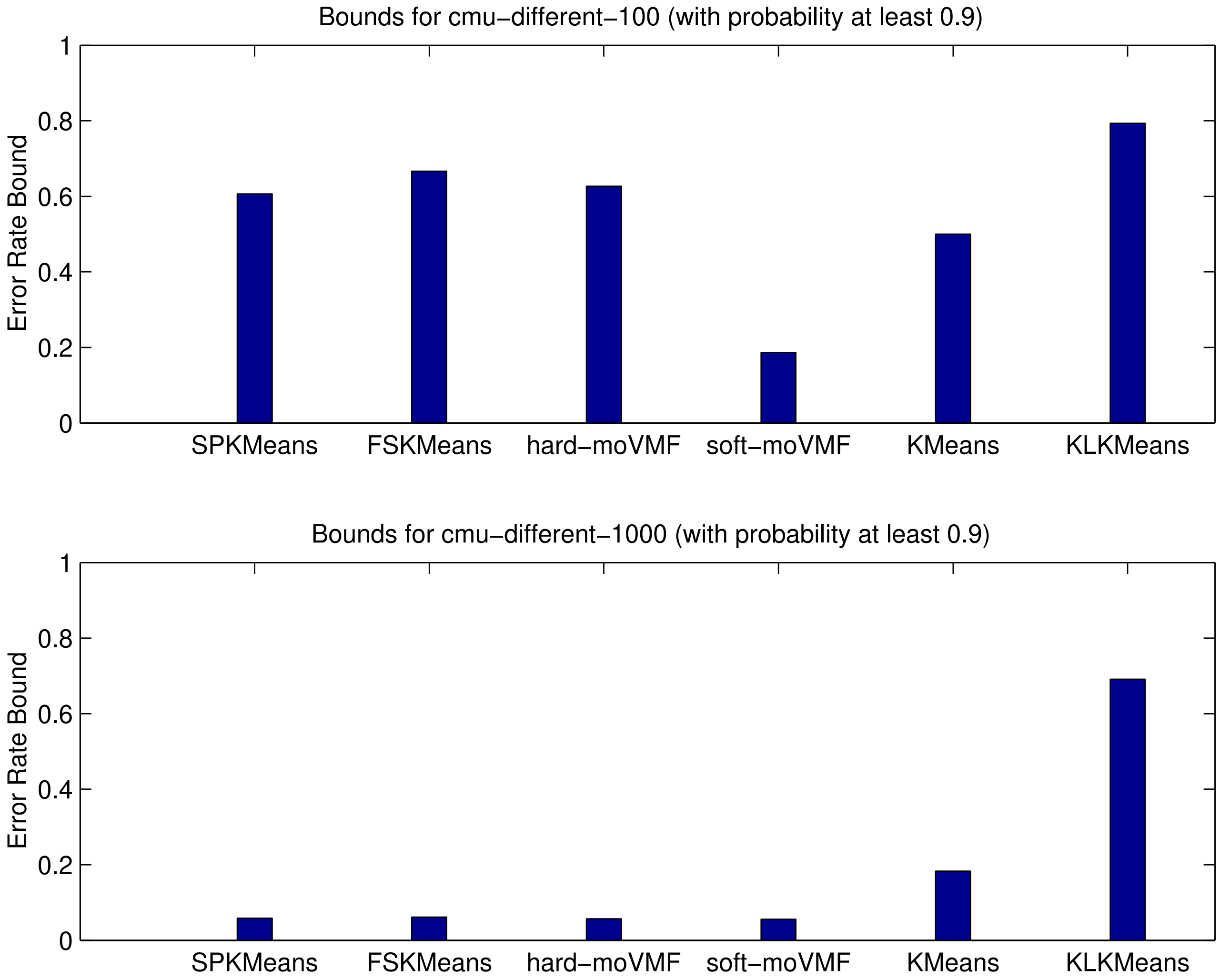
r random initializations $\Rightarrow P(\sigma) = \frac{2^{-\sigma}}{r}$

k unknown $\Rightarrow P(\sigma) = \frac{2^{-k}}{k(k+1)r}$

t possible clustering algorithms $\Rightarrow P(\sigma) = \frac{2^{-k}}{k(k+1)tr}$



$$P(\sigma) = \frac{2^{-k}}{r}$$



$$P(\sigma) = \frac{2^{-k}}{k(k+1)r}$$

Next Time: Transductive bound implies Inductive bound.