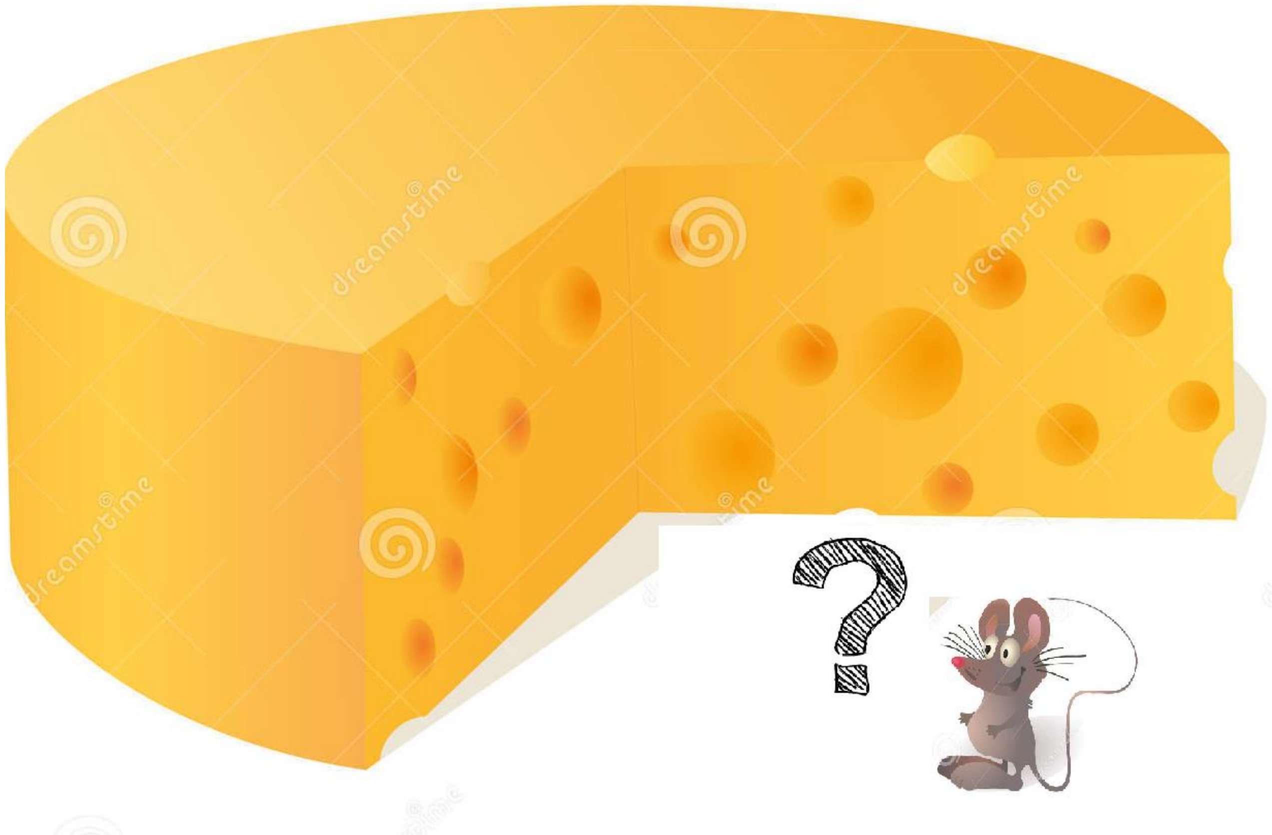
Big Data Class



LECTURER: DAN FELDMAN

TEACHING ASSISTANTS:

IBRAHIM JUBRAN

ALAA MAALOUF

אוניברסיטת חיפה
University of Haifa
جامعة حيفا

Probability

Hoeffding's bound
(Simplified version – from Young'95 paper)

- For n independent random variables $X_1, \dots, X_n$ where $X_i \in [0,1]$, with $E(X_i) \leq \mu_i$
- Let $X = \sum_{i=1}^{n} \frac{X_i}{n}$ and $\mu = \sum_{i=1}^{n} \mu_i$.
- Then: $\Pr[X \geq \mu + n\epsilon] < \frac{1}{e^{2n\epsilon^2}}$

Hoeffding's bound - Proof
(Simplified version – from Young'95 paper)

- Let $\alpha = e^{4\epsilon} - 1$.

Hoeffding's bound - Proof
(Simplified version – from Young'95 paper)

- Let $\alpha = e^{4\epsilon} - 1$.

$$\Pr[\, \sum_{i=1}^{n} X_i \geq \mu + n\epsilon \,]$$
$$= \Pr\left[\, \prod_{i=1}^{n} \frac{(1+\alpha)^{X_i}}{(1+\alpha)^{\mu_i+\epsilon}} \geq 1 \,\right]$$

Hoeffding's bound - Proof
(Simplified version – from Young'95 paper)

- Let $\alpha = e^{4\epsilon} - 1$.

$$\Pr\left[\sum_{i=1}^{n} X_i \geq \mu + n\epsilon\right]$$

$$= \Pr\left[\prod_{i=1}^{n} \frac{(1+\alpha)^{X_i}}{(1+\alpha)^{\mu_i+\epsilon}} \geq 1\right]$$

$$\leq \mathrm{E}\left[\prod_{i=1}^{n} \frac{1+\alpha X_i}{(1+\alpha)^{\mu_i+\epsilon}}\right]$$

Follows from:

1. For $0 \leq z \leq 1$, $(1+\alpha)^z \leq 1 + \alpha z$
2. Markov's inequality : $\Pr[X \geq a] \leq \frac{E(X)}{a}$

Hoeffding's bound - Proof
(Simplified version – from Young'95 paper)

- Let $\alpha = e^{4\epsilon} - 1$.

$$\Pr[\, \sum_{i=1}^{n} X_i \geq \mu + n\epsilon \,]$$

$$= \Pr\left[\prod_{i=1}^{n} \frac{(1+\alpha)^{X_i}}{(1+\alpha)^{\mu_i+\epsilon}} \geq 1 \right]$$

$$\leq \mathrm{E}\left[\prod_{i=1}^{n} \frac{1+\alpha X_i}{(1+\alpha)^{\mu_i+\epsilon}} \right]$$

$$= \prod_{i=1}^{n} \frac{1+\alpha E(X_i)}{(1+\alpha)^{\mu_i+\epsilon}}$$

Follows from:

1. For $0 \leq z \leq 1$, $(1+\alpha)^z \leq 1 + \alpha z$
2. Markov's inequality : $\Pr[X \geq a] \leq \frac{E(X)}{a}$

Hoeffding's bound - Proof
(Simplified version – from Young'95 paper)

- Let $\alpha = e^{4\epsilon} - 1$.

$$\Pr\left[\sum_{i=1}^{n} X_i \geq \mu + n\epsilon\right]$$

$$= \Pr\left[\prod_{i=1}^{n} \frac{(1+\alpha)^{X_i}}{(1+\alpha)^{\mu_i+\epsilon}} \geq 1\right]$$

$$\leq \mathrm{E}\left[\prod_{i=1}^{n} \frac{1+\alpha X_i}{(1+\alpha)^{\mu_i+\epsilon}}\right]$$

$$= \prod_{i=1}^{n} \frac{1+\alpha E(X_i)}{(1+\alpha)^{\mu_i+\epsilon}}$$

$$\leq \prod_{i=1}^{n} \frac{1+\alpha \mu_i}{(1+\alpha)^{\mu_i+\epsilon}}$$

Follows from:

1. For $0 \leq z \leq 1$, $(1+\alpha)^z \leq 1 + \alpha z$
2. Markov's inequality : $\Pr[X \geq a] \leq \frac{E(X)}{a}$

Hoeffding's bound - Proof
(Simplified version – from Young'95 paper)

- Let $\alpha = e^{4\epsilon} - 1$.

$$\Pr[\, \textstyle\sum_{i=1}^{n} X_i \geq \mu + n\epsilon \,]$$

$$= \Pr\left[\prod_{i=1}^{n} \frac{(1+\alpha)^{X_i}}{(1+\alpha)^{\mu_i+\epsilon}} \geq 1 \right]$$

$$\leq \mathrm{E}\left[\prod_{i=1}^{n} \frac{1+\alpha X_i}{(1+\alpha)^{\mu_i+\epsilon}} \right]$$

$$= \prod_{i=1}^{n} \frac{1+\alpha E(X_i)}{(1+\alpha)^{\mu_i+\epsilon}}$$

$$\leq \prod_{i=1}^{n} \frac{1+\alpha \mu_i}{(1+\alpha)^{\mu_i+\epsilon}}$$

$$< \prod_{i=1}^{n} \frac{1}{e^{2\epsilon^2}}$$

Follows from:

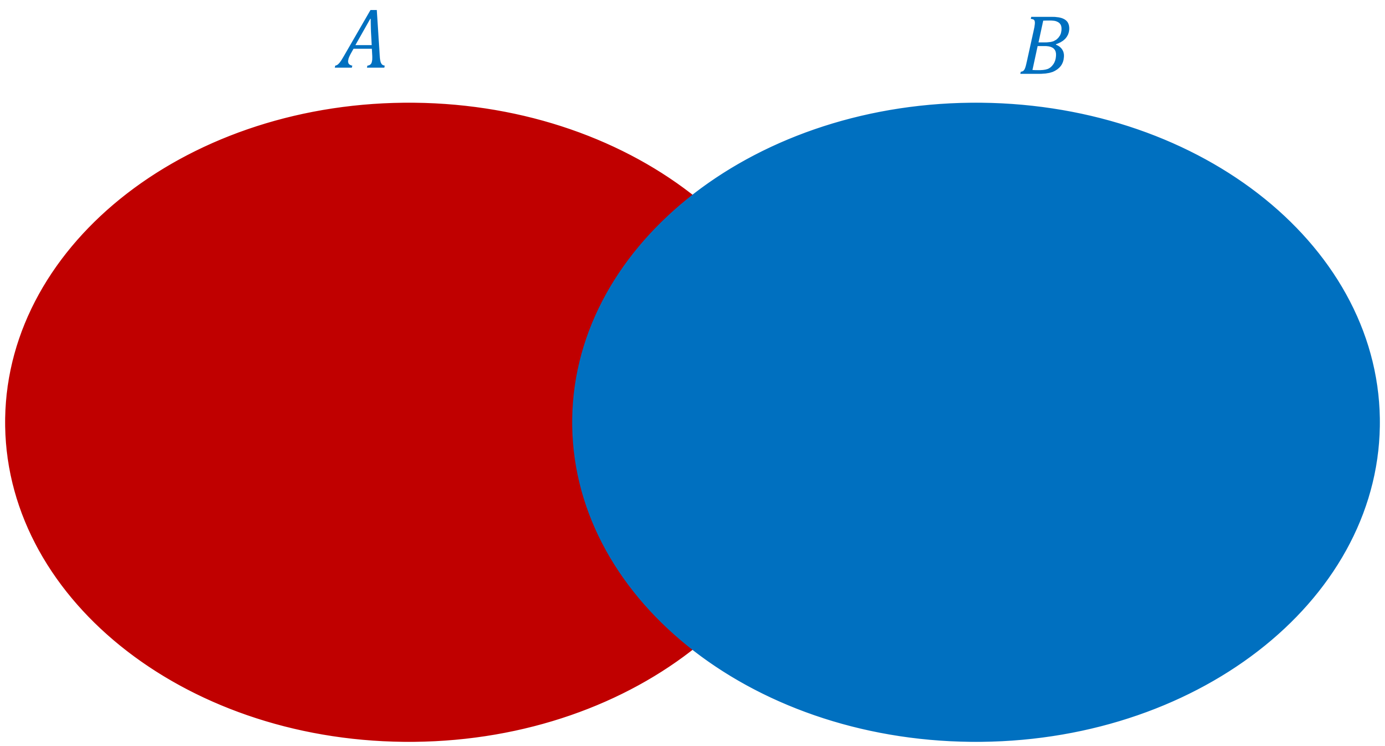
1. For $0 \leq z \leq 1$, $(1+\alpha)^z \leq 1 + \alpha z$
2. Markov's inequality : $\Pr[X \geq a] \leq \frac{E(X)}{a}$

For $\epsilon > 0$, $\alpha = e^{4\epsilon} - 1$, $z \geq 0$:

$$1 + \alpha z < \frac{(1+\alpha)^{z+\epsilon}}{e^{2\epsilon^2}}$$

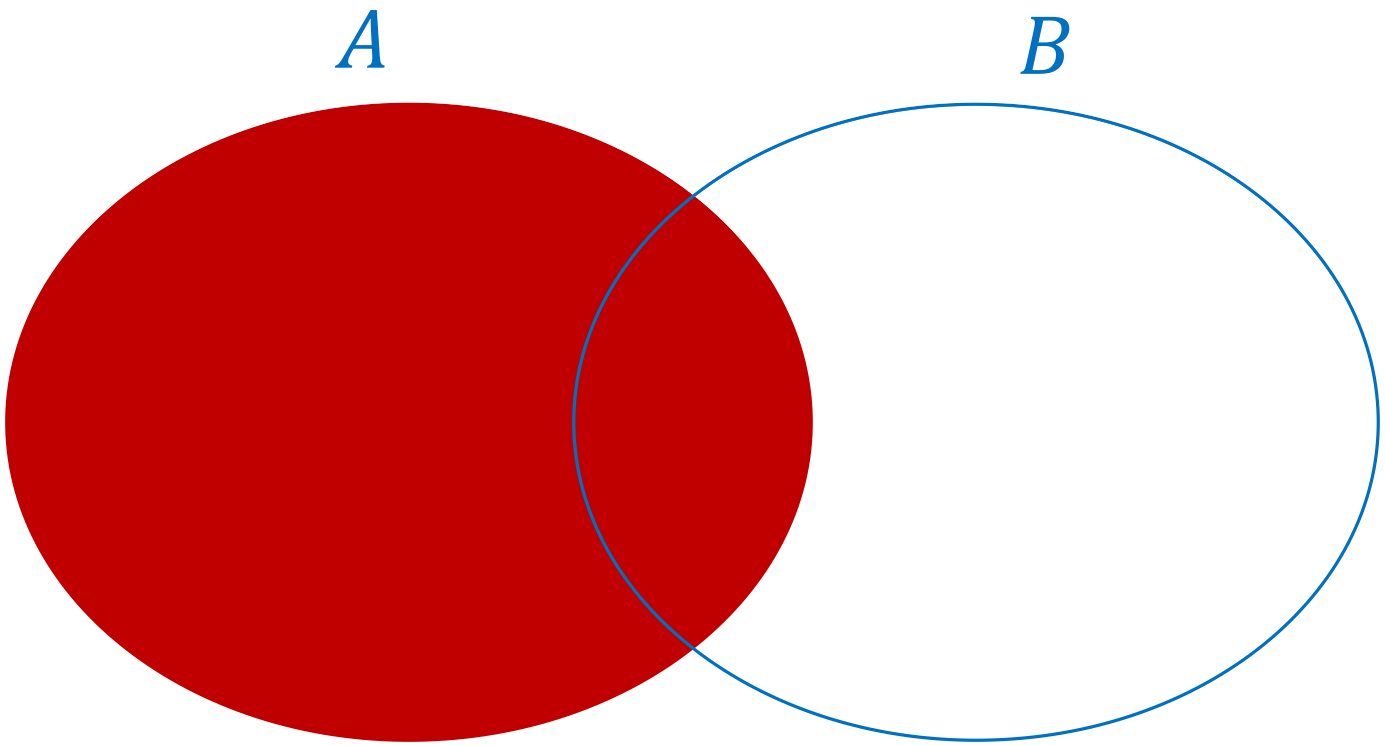
Union Bound (Boole's inequality)

$A \cup B =$



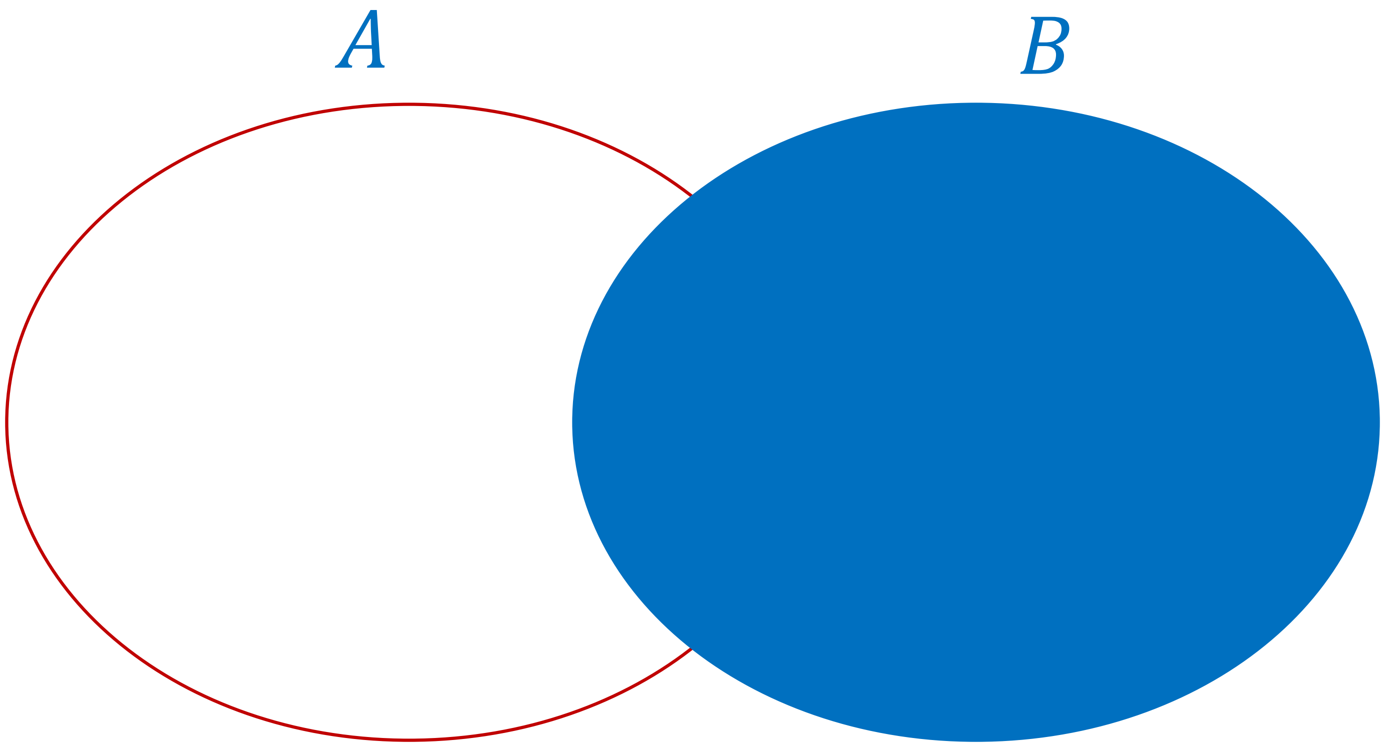
Union Bound (Boole's inequality)

$A \cup B = A$



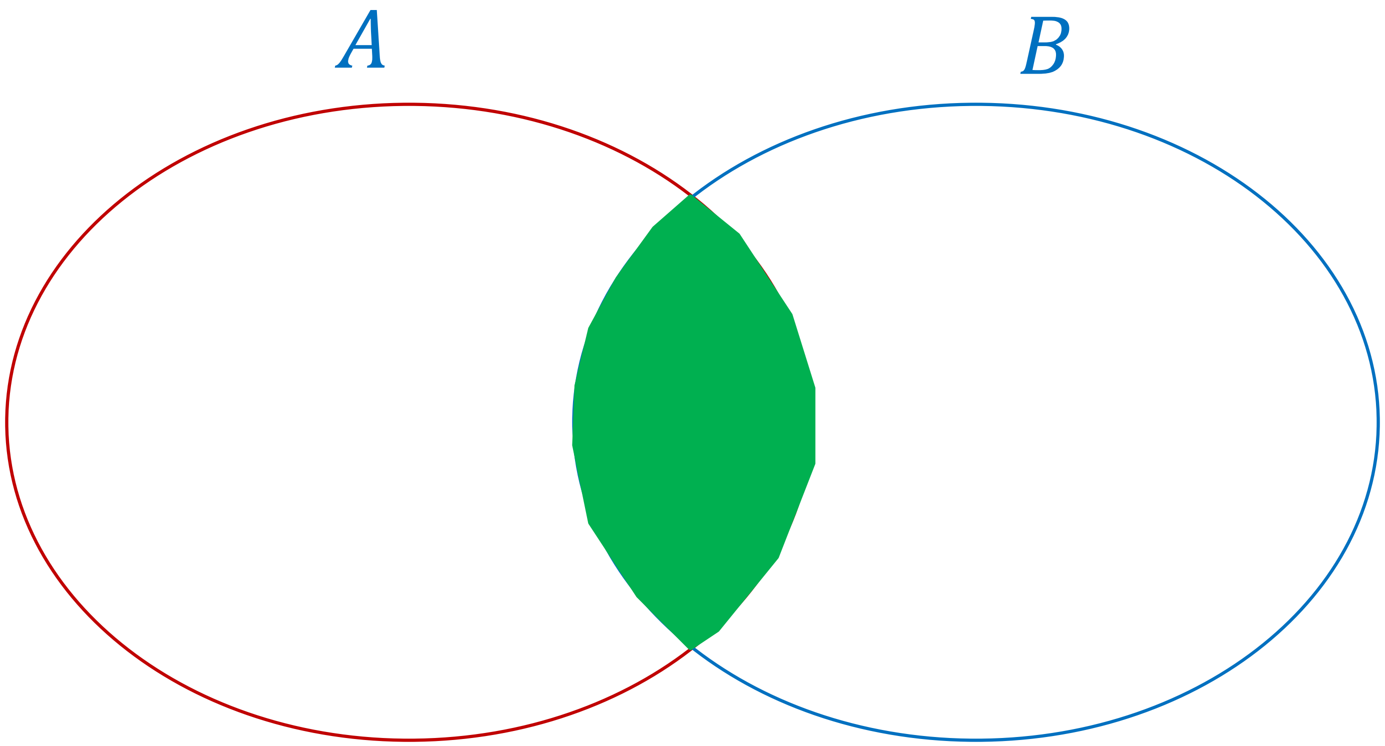
Union Bound (Boole's inequality)

$A \cup B = A + B$



Union Bound (Boole's inequality)

$A \cup B = A + B - A \cap B$



Union Bound (Boole's inequality)

$$\rightarrow \; P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Union Bound (Boole's inequality)

$\rightarrow \; P(A \cup B) = P(A) + P(B) - P(A \cap B)$

→ For any events $A_1, A_2, \dots, A_n$:

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i)$$

Union Bound (Boole's inequality)

$\rightarrow \; P(A \cup B) = P(A) + P(B) - P(A \cap B)$

→ For any events $A_1, A_2, \dots, A_n$:

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i)$$

$$P\left(\bigcup_{i=1}^{n} A_i\right) \geq \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P\left(A_i \cap A_j\right)$$

Union Bound (Boole's inequality)

$\rightarrow \; P(A \cup B) = P(A) + P(B) - P(A \cap B)$

→ For any events $A_1, A_2, \ldots, A_n$:

$$\boxed{P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i)}$$

$$P\left(\bigcup_{i=1}^{n} A_i\right) \geq \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j)$$

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k)$$

⋮

Hoeffding's bound → ϵ-sample

Pick a random sample S of F, it holds that:

$$\underbrace{Pr_1}_{\text{Probability of failure for } range_1} = Pr\underbrace{\left(\left|\frac{|F \cap range_1|}{|F|} - \frac{|S \cap range_1|}{|S|}\right| > \epsilon\right)}_{S \text{ is an } \epsilon\text{-sample}}$$

Hoeffding's bound → ϵ-sample

Pick a random sample S of F, it holds that:

$$Pr_1 = Pr\left(\left|\frac{|F \cap range_1|}{|F|} - \frac{|S \cap range_1|}{|S|}\right| > \epsilon\right) < \frac{1}{e^{2|S|\epsilon^2}}$$

Probability of failure for $range_1$

Hoeffding

Hoeffding's bound → ϵ-sample

Pick a random sample S of F, it holds that:

$$\underset{\text{Probability of failure for } range_1}{Pr_1} = Pr\left(\left|\frac{|F \cap range_1|}{|F|} - \frac{|S \cap range_1|}{|S|}\right| > \epsilon\right) \underset{\text{Hoeffding}}{<} \frac{1}{e^{2|S|\epsilon^2}}$$

Probability of failure for $range_1$

Hoeffding

$$\mathrm{Pr_{bad}} = Pr_1 \cup Pr_2 \cup \cdots \cup Pr_m \underset{\text{Union Bound}}{\leq} m \cdot \frac{1}{e^{2|S|\epsilon^2}}$$

Union Bound

Hoeffding's bound → ϵ-sample

The probability for failure should be small:

$$\Pr_{\text{bad}} \leq m \cdot \frac{1}{e^{2|S|\epsilon^2}} \leq \delta$$

$$\to \frac{m}{\delta} \leq e^{2|S|\epsilon^2}$$

$$\to \boxed{|S| \geq \frac{1}{\epsilon^2}\left(\log m + \log\frac{1}{\delta}\right)}$$

Hoeffding's bound → ϵ-sample

The probability for failure should be small:

$$\Pr_{\text{bad}} \leq m \cdot \frac{1}{e^{2|S|\epsilon^2}} \leq \delta$$

$$\rightarrow \frac{m}{\delta} \leq e^{2|S|\epsilon^2}$$

$$\rightarrow \boxed{|S| \geq \frac{1}{\epsilon^2}\left(\log m + \log \frac{1}{\delta}\right)}$$

Might be infinite!

Handling $m \to \infty$

Example:

Q = circles in $R^d \to range(F, q, r) = \{f \in F$ | f inside the circle with center q and radius $r\}$

$\to$ number of different circles: $|Q| = m = \infty$.

However, the number of different equivalence classes is $\underline{n^{O(d)}}$ since: A sphere in R^d is determined by $d + 1$ points.

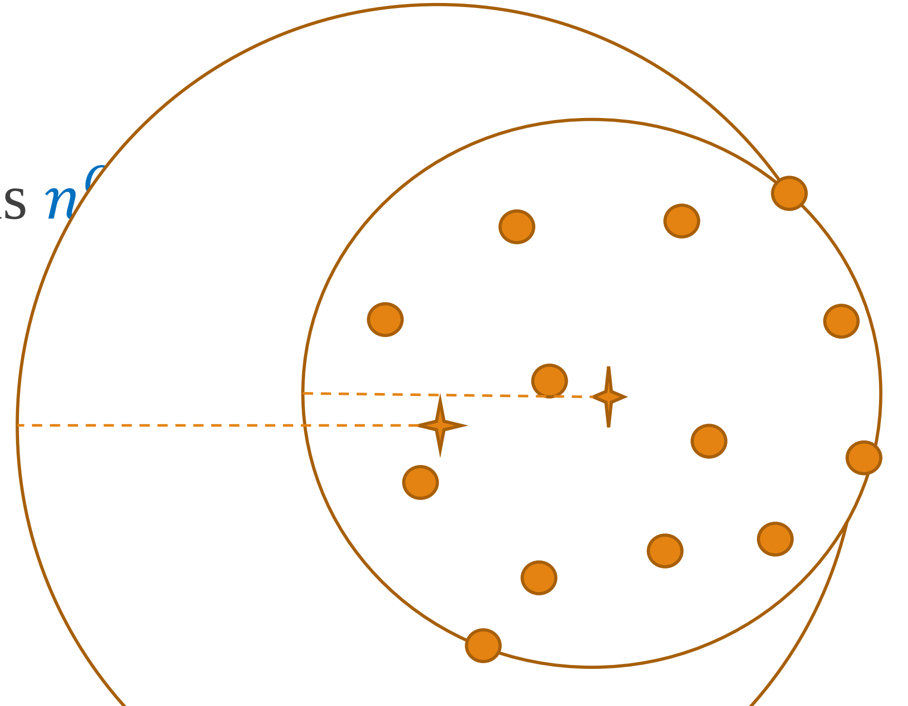
Handling $m \to \infty$

Example:

Q = circles in R^d $\to range(F, q, r) = \{f \in F$ | f inside the circle with center q and radius $r\}$

$\to$ number of different circles: $|Q| = m = \infty$.

However, the number of different equivalence classes is n
R^d is determined by $d + 1$ points.



Handling $m \to \infty$

Example:

Q = circles in $R^d \to range(F, q, r) = \{f \in F$ | f inside the circle with center q and radius $r\}$

$\to$ number of different circles: $|Q| = m = \infty$.

However, the number of different equivalence classes is $\underline{n^{O(d)}}$ since: A sphere in R^d is determined by $d + 1$ points.

$$\to |S| \geq \frac{1}{\epsilon^2}\left(\log n^d + \log\frac{1}{\delta}\right) = \frac{1}{\epsilon^2}\left(d \log n + \log\frac{1}{\delta}\right)$$

VC-dimension

Definition: (Range Space)

A range space is a pair $(F, ranges)$ where F is a set, called ground set and $ranges$ is a family (set) of subsets of F.

Definition: (VC-dimension)

The VC-dimension of a range space $(F, ranges)$ is the size $|G|$ of the largest subset $G \subseteq F$ such that

$$|\{ G \cap range \mid range \in ranges \}| \leq 2^{|G|}$$

VC-dimension

Theorem:

Let $f_1, \ldots, f_m$ be real polynomials in $d \leq m$ variables, each of constant degree. Then the number of sign sequences $\left(sign\left(f_1(x)\right), \ldots, sign\left(f_m(x)\right)\right), x \in R^d$, that consist of 1 and -1 is at most $O\left(\left(\frac{m}{d}\right)^d\right)$.

VC-dimension

Theorem:

Let $f_1, \ldots, f_m$ be real polynomials in $d \leq m$ variables, each of constant degree. Then the number of sign sequences $\left(sign\left(f_1(x)\right), \ldots, sign\left(f_m(x)\right)\right), x \in R^d$, that consist of 1 and -1 is at most $O\left(\left(\frac{em}{d}\right)^d\right)$.

Corollary:

If $m \geq O(d)$, then the number of distinct sequences as in the above theorem is less than 2^m.

VC-dimension

Theorem:

Let $f_1, \dots, f_m$ be real polynomials in $d \le m$ variables, each of constant degree. Then the number of sign sequences $\left(sign\left(f_1(x)\right), \dots, sign\left(f_m(x)\right)\right), x \in R^d$, that consist of 1 and -1 is at most $O\left(\left(\frac{em}{d}\right)^d\right)$.

Corollary:

If $m \ge O(d)$, then the number of distinct sequences as in the above theorem is less than 2^m.

Proof: for which values of m it holds that $\left(\frac{em}{d}\right)^d \le 2^m$?

VC-dimension

Theorem:

Let $f_1, \ldots, f_m$ be real polynomials in $d \leq m$ variables, each of constant degree. Then the number of sign sequences $\left(sign\left(f_1(x)\right), \ldots, sign\left(f_m(x)\right)\right), x \in R^d$, that consist of 1 and -1 is at most $O\left(\left(\frac{em}{d}\right)^d\right)$.

Corollary:

If $m \geq O(d)$, then the number of distinct sequences as in the above theorem is less than 2^m.

Proof:

$$\left(\frac{em}{d}\right)^d \leq 2^m$$

VC-dimension

Theorem:

Let $f_1, \ldots, f_m$ be real polynomials in $d \leq m$ variables, each of constant degree. Then the number of sign sequences $\left(sign\left(f_1(x)\right), \ldots, sign\left(f_m(x)\right)\right), x \in R^d$, that consist of 1 and -1 is at most $O\left(\left(\frac{em}{d}\right)^d\right)$.

Corollary:

If $m \geq O(d)$, then the number of distinct sequences as in the above theorem is less than 2^m.

Proof:

$\wedge\frac{1}{d}$

$$\left(\frac{em}{d}\right)^d \leq 2^m \rightarrow \frac{em}{d} \leq 2^{\frac{m}{d}}$$

VC-dimension

Theorem:

Let $f_1, \dots, f_m$ be real polynomials in $d \le m$ variables, each of constant degree. Then the number of sign sequences $\left(sign\left(f_1(x)\right), \dots, sign\left(f_m(x)\right)\right), x \in R^d$, that consist of 1 and -1 is at most $O\left(\left(\frac{em}{d}\right)^d\right)$.

Corollary:

If $m \ge O(d)$, then the number of distinct sequences as in the above theorem is less than 2^m.

$ex \le 2^x$ for $x \ge 4$

Proof:

$$\left(\frac{em}{d}\right)^d \le 2^m \rightarrow \frac{em}{d} \le 2^{\frac{m}{d}} \rightarrow \frac{m}{d} \ge 4$$

VC-dimension

Theorem:

Let $f_1, \dots, f_m$ be real polynomials in $d \le m$ variables, each of constant degree. Then the number of sign sequences $\left(sign\left(f_1(x)\right), \dots, sign\left(f_m(x)\right)\right), x \in R^d$, that consist of 1 and -1 is at most $O\left(\left(\frac{em}{d}\right)^d\right)$.

Corollary:

If $m \ge O(d)$, then the number of distinct sequences as in the above theorem is less than 2^m.

Proof:

$$\left(\frac{em}{d}\right)^d \le 2^m \rightarrow \frac{em}{d} \le 2^{\frac{m}{d}} \rightarrow \frac{m}{d} \ge 4 \rightarrow m = O(d) \quad \blacksquare$$

VC-dimension

Let $Q = \{q_1, \dots, q_k\} \subseteq R^d$ and $R = \{r_1, \dots, r_k\} \subseteq R$.

Then:

$$range(P, Q, R) = \left\{ p \in P \mid \vee_i \left(dist^2(p, q_i) \leq r_i^2 \right) \right\}$$

Consider the following polynomials:

$$Poly(P, Q, R) = \left\{ \left\| p_i - q_j \right\|^2 - r_j^2 \;\middle|\; i \in [n], j \in [k] \right\}$$

$\rightarrow$ $|Poly(P, Q, R)| = nk$ such polynomials in dk variables.

VC-dimension

Let $Q_1, Q_2 \subseteq R^d$ and $R_1, R_2 \subseteq R$.

Lemma:

If $Poly(P, Q_1, R_1)$ and $Poly(P, Q_2, R_2)$ have the same sign sequence for the nk polynomials, then

$$range(P, Q_1, R_1) = range(P, Q_2, R_2)$$

VC-dimension

Let $Q_1, Q_2 \subseteq R^d$ and $R_1, R_2 \subseteq R$.

Lemma:

If $Poly(P, Q_1, R_1)$ and $Poly(P, Q_2, R_2)$ have the same sign sequence for the nk polynomials, then

$$range(P, Q_1, R_1) = range(P, Q_2, R_2)$$

Proof:

Let $p \in range(P, Q_1, R_1)$

$\rightarrow$ exists $j \in k$ such that $\|p - \underbrace{q_{j1}}_{Q_1}\|^2 - \underbrace{r_{j1}^2}_{R_1} \leq 0$

VC-dimension

Let $Q_1, Q_2 \subseteq R^d$ and $R_1, R_2 \subseteq R$.

<u>Lemma:</u>

If $Poly(P, Q_1, R_1)$ and $Poly(P, Q_2, R_2)$ have the same sign sequence for the nk polynomials, then

$$range(P, Q_1, R_1) = range(P, Q_2, R_2)$$

<u>Proof:</u>

Let $p \in range(P, Q_1, R_1)$

$\rightarrow$ exists $\mathrm{j} \in k$ such that $\left\|p - \overset{\overset{Q_1}{\uparrow}}{q_{j1}}\right\|^2 - \overset{\overset{R_1}{\uparrow}}{r_{j1}^2} \leq 0 \rightarrow \left\|p - \overset{\overset{Q_2}{\uparrow}}{q_{j2}}\right\|^2 - \overset{\overset{R_2}{\uparrow}}{r_{j2}^2} \leq 0$

VC-dimension

Let $Q_1, Q_2 \subseteq R^d$ and $R_1, R_2 \subseteq R$.

Lemma:

If $Poly(P, Q_1, R_1)$ and $Poly(P, Q_2, R_2)$ have the same sign sequence for the nk polynomials, then

$$range(P, Q_1, R_1) = range(P, Q_2, R_2)$$

Proof:

Let $p \in range(P, Q_1, R_1)$

$\rightarrow$ exists j $\in k$ such that $\|p - \overset{\overset{Q_1}{\uparrow}}{q_{j1}}\|^2 - \overset{\overset{R_1}{\uparrow}}{r_{j1}^2} \leq 0 \rightarrow \|p - \overset{\overset{Q_2}{\uparrow}}{q_{j2}}\|^2 - \overset{\overset{R_2}{\uparrow}}{r_{j2}^2} \leq 0$

$\rightarrow p \in range(P, Q_2, R_2)$

$\blacksquare$

VC-dimension

Conclusion:

#different ranges $\leq$ #different sign sequences $\leq 2^m$

Last Lemma

If $m > O(d)$