Εκμάθηση Διακριτών Κατανομών

Έχουμε n στοιχεία

$[n] = \{1, 2, \dots, n\}$

Μια κατανομή $p \in [0,1]^n$ με $\sum_{i=1}^{n} p_i = 1$

Ο χώρος όλων των κατανομών

$$D = \{ q \in [0,1]^n : \sum_{i=1}^{n} q_i = 1 \}$$

Total Variation Distance

(ή ℓ_1 - distance)

$$d_{TV}(p, q) = \frac{1}{2} \sum_{i=1}^{n} |p_i - q_i| = \frac{1}{2} \|p - q\|_1$$

$n=2$ < κορώνα 1
γράμματα 2

$$d_{TV}\left((p, 1-p), (q, 1-q)\right)$$

$$= \frac{1}{2}\left(|p-q| + |1-p-(1-q)|\right)$$

$$= |p-q|$$

Πόσα δείγματα χρειαζόμαστε από μια κατανομή p για να μάθουμε $\hat{p}$ με $d_{TV}(\hat{p}, p) \leq \varepsilon$ με πιθανότητα $1-\delta$

Εκτιμητής: Με m δείγματα $X_1, \dots, X_m$

$$\hat{p}_i = \frac{\sum_{j=1}^{m} 1[X_j = i]}{m}$$

$$E[\, d_{TV}(\hat{p}, p)] \quad = ?$$

$$E[\hat{p}_i] = E\left[\frac{\sum_{j=1}^{m} 1\{x_j = i\}}{m}\right] = \frac{1}{m} m p_i$$

$$= p_i$$

$$\mathrm{Var}[\hat{p}_i] = \frac{1}{m^2} m \, \mathrm{Var}[1\{x_j = i\}]$$

$$= \frac{1}{m} \quad p_i(1 - p_i) \leq \frac{1}{m} p_i$$

$$E[|\hat{p}_i - p_i|] \leq \sqrt{E[(\hat{p}_i - p_i)^2]}$$

$$= \sqrt{\mathrm{Var}[\hat{p}_i]} \leq \sqrt{\frac{p_i}{m}}$$

$$E[\, d_{TV}(\hat{p}, p)]$$

$$= \frac{1}{2} \sum_{i=1}^{n} E\left[|\hat{p}_i - p_i| \right]$$

$$\leq \frac{\sum_{i=1}^{n} \sqrt{p_i}}{2\sqrt{m}} \leq \frac{\sqrt{n}}{2\sqrt{m}}$$

Αν $\boxed{m = 4\frac{n}{\varepsilon^2}}$ τότε

$$E\left[d_{TV}(\hat{p}, p) \right] \leq \frac{\varepsilon}{4}$$

άρα

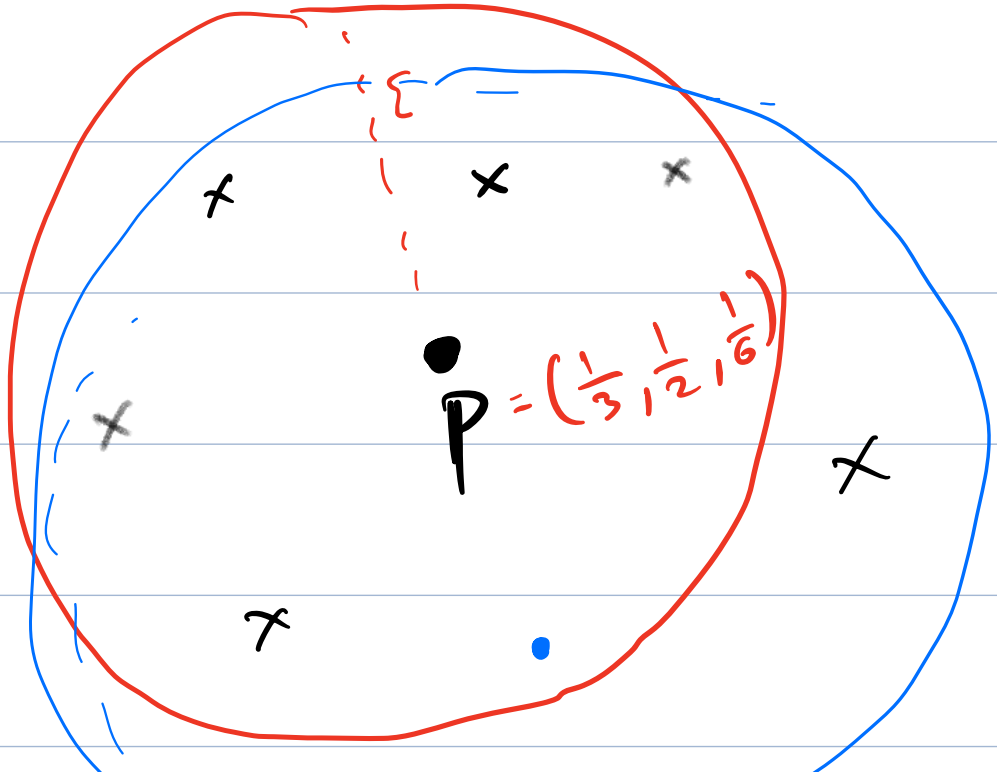
$$\Pr\left[d_{TV}(\hat{p}, p) > \varepsilon \right] \leq \frac{1}{4}$$

από Markov

Συνολικά με $m = \Theta\left(\frac{n}{\varepsilon^2}\right)$ δείγματα βρίσκουμε $\hat{p}$ τέτοιο ώστε

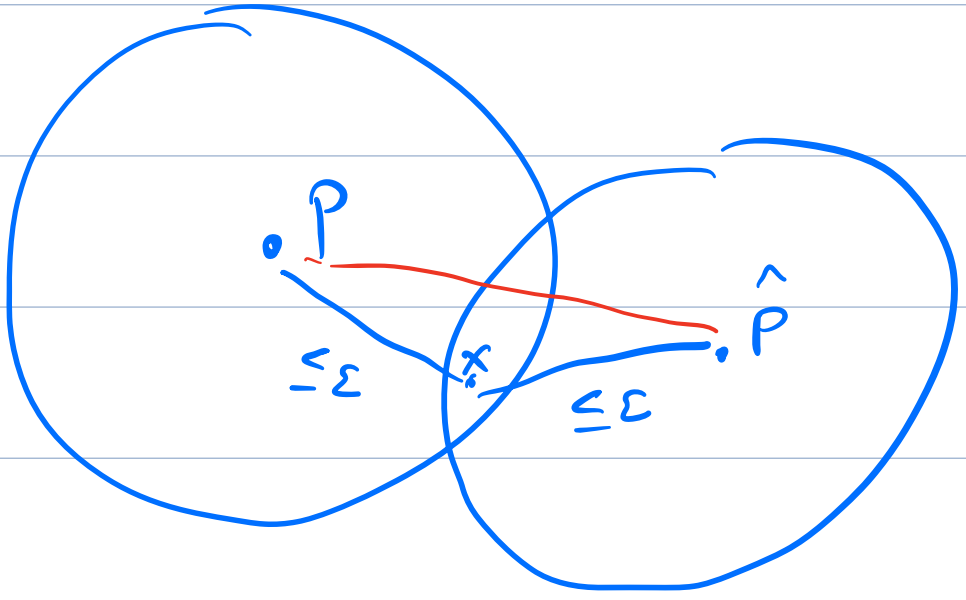
$$d_{TV}(\hat{p}, p) \leq \varepsilon \quad \text{με πιθαν. } \frac{3}{4}$$

Πως μπορούμε να αυξήσουμε την πιθανότητα;



Θα βρώ $\hat{p}$ η οποία

είναι πιο κοντά στις περιστροφές

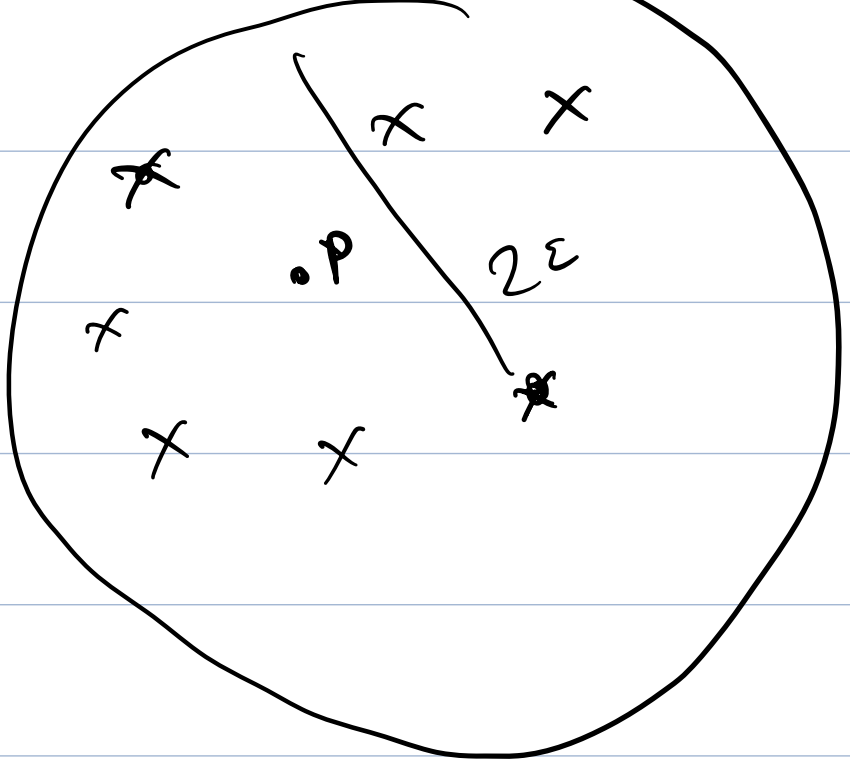


με $\Theta(\log(1/\delta))$ επαναλήψεις

με πιθανότητα δ τουλάχιστον

πάνω από τα μισά:

$$d_H(\hat{p}, p) \leq 2\varepsilon$$



$$\Theta\left(\frac{n + \log(1/\delta)}{\varepsilon^2}\right) \quad \text{δείγματα}$$

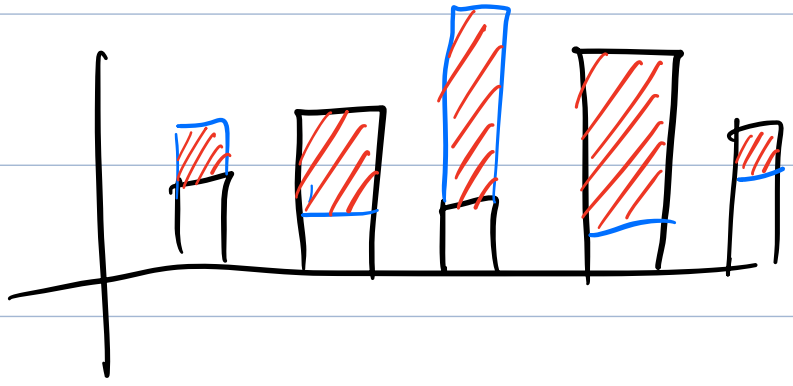
αρκούν για να μάθω σε TV-distance ε μία πιθανότητα k.δ.

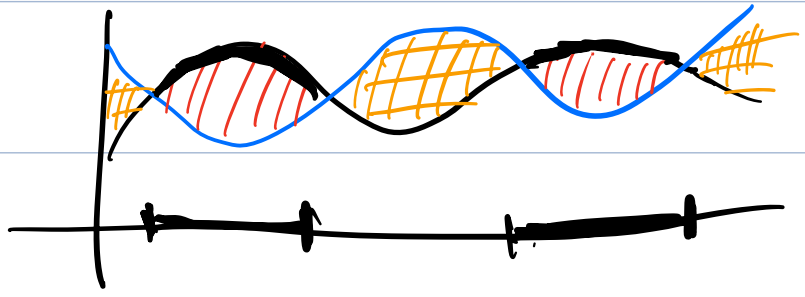
Μετρικές στο χώρο των κατανομών

- TV - distance p, q

$$d_{TV}(p,q) = \frac{1}{2}\sum |p_i - q_i| = \sum_{i=1}^{n} (p_i - q_i)_+$$

$$= \sum_{i=1}^{n} (q_i - p_i)_+$$





$$d_{TV}(p,q) = \max_{S \subseteq [n]} |p(S) - q(S)|$$

$$\left| \Pr_{i \sim p}[i \in S] - \Pr_{i \sim q}[i \in S] \right|$$

Για να μάθω $\hat{p}$ τ.ω. $d_{TV}(\hat{p}, p) \leq \varepsilon$ με πιθανότητα $1-\delta$ χρειάζομαι $O\left(\frac{n + \log(1/\delta)}{\varepsilon^2}\right)$ δείγματα

$$\hat{p}_i = \frac{\sum_{j=1}^{m} 1\{x_j = i\}}{m}$$

Φτιάχνω ένα $S \subseteq [n]$

$$\Pr\left[\left| \hat{p}(S) - p(S) \right| \geq \varepsilon \right] \leq e^{-\Omega(\varepsilon^2 m)}$$

$\sum_{i \in S} p_i$

Για να ισχυει $d_{TV}(p, \hat{p}) \leq \varepsilon$

πρεπει $\forall S \subseteq [n]$

$$|p(S) - \hat{p}(S)| \leq \varepsilon$$

$$\Pr[\exists S \in [n] : |p(S) - \hat{p}(S)| > \varepsilon]$$

$$\leq \sum_{S \subseteq [n]} \Pr[|p(S) - \hat{p}(S)| > \varepsilon]$$

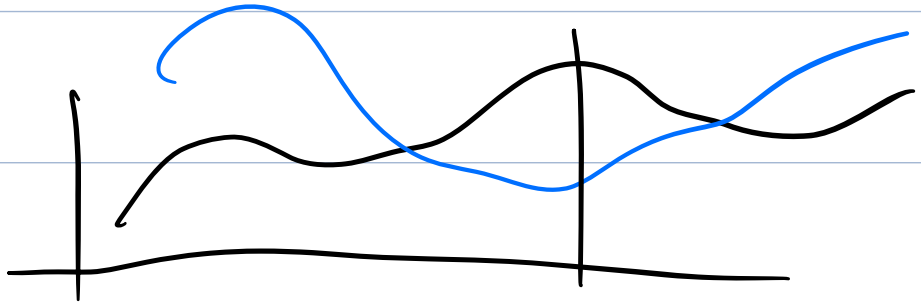
(Union bound)

$$\leq 2^n \, e^{-\Omega(\varepsilon^2 m)}$$

$$\leq \delta$$

Αρα $m = O\left(\frac{n + \log(1/\delta)}{\varepsilon^2}\right)$

- Kolmogoron Distance



$$d_K(p, q) = \max_t \left| \Pr_{x \sim p}[x \leq t] - \Pr_{x \sim q}[x \leq t] \right|$$

- KL-divergence

$$D_{KL}(p \| q) = \sum_i p_i \log \frac{p_i}{q_i}$$

$$= \sum p_i \log p_i - \sum p_i \log q_i$$

$$= C - \underset{i \sim p}{E}[\log(q_i)]$$

$$D_{KL}(p \| q) \geq 0$$

Pinsker's Inequality

$$d_{TV}(p, q) \leq \sqrt{\frac{1}{2} D_{KL}(p \| q)}$$

Εκτίμηση πιθανότητας κέρματος

Lower Bound

Χρειάζεται $1/\varepsilon^2$;

Έστω $p = Ber(1/2)$

$q = Ber(1/2 + \varepsilon)$

Θ.δ.ο. για πιθανότητα λάθους $1/4$ χρειάζονται $\Omega(1/\varepsilon^2)$ δείγματα

Le Cam's method

Έστω 2 κατανομές p, q σε ένα χώρο X

Στόχος: Να βρούμε test

$$T : X \to \{p, q\}$$

με το error probability

$$\max \left\{ \Pr_{x \sim p} [T(x) = q], \Pr_{x \sim q} [T(x) = p] \right\}$$

Λήμμα: $\text{Error Prob} \geq \frac{1}{2}(1 - d_{TV}(p, q))$

Εφαρμογή αποδείξης

$$P = Ber(1/2)^{\otimes m}$$

$$Q = Ber(1/2 + \varepsilon)^{\otimes m}$$

$$\text{Prob of Error} \geq \frac{1}{2}\left(1 - d_{TV}(P, Q)\right)$$

[Claim: $d_{TV}(P, Q) \leq O(\sqrt{m}\,\varepsilon)$]

$$\geq \frac{1}{2}\left(1 - O(\sqrt{m}\,\varepsilon)\right)$$

$$\geq \frac{1}{4}$$ εκτός αν

$$m = \Omega\left(\frac{1}{\varepsilon^2}\right)$$

[Claim : $d_{TV}(P, Q) \le O(\sqrt{m}\,\varepsilon)$]

$$d_{KL}(P \| Q) = \sum_{x \in \{0,1\}^m} P_x \log \frac{P_x}{Q_x}$$

$$= \sum_{x \in \{0,1\}^m} \prod_{i=1}^{m} p(x_i) \log \frac{\prod_{i=1}^{m} p(x_i)}{\prod_{i=1}^{m} q(x_i)}$$

$$= \sum_{i=1}^{m} \underbrace{\sum_{x_{-i} \in \{0,1\}^{m-1}} \prod_{j \neq i} p(x_j)}_{1} \underbrace{\sum_{x_i \in \{0,1\}} p(x_i) \log \frac{p(x_i)}{q(x_i)}}_{d_{KL}(p, q)}$$

$$= m\, d_{KL}(p, q)$$

$$= m\, d_{KL}\left(\mathrm{Ber}\left(\tfrac{1}{2}\right), \mathrm{Ber}\left(\tfrac{1}{2} + \varepsilon\right)\right)$$

$$d_{KL}\left(\text{Ber}(1/2), \text{Ber}\left(\tfrac{1}{2}+\varepsilon\right)\right)$$

$$= \frac{1}{2}\log\frac{1/2}{1/2+\varepsilon} + \frac{1}{2}\log\frac{1/2}{1/2-\varepsilon}$$

$$= \frac{1}{2}\log\frac{1}{1+2\varepsilon} + \frac{1}{2}\log\frac{1}{1-2\varepsilon}$$

$$= \frac{1}{2}\log\frac{1}{(1+2\varepsilon)(1-2\varepsilon)} = \frac{1}{2}\log\frac{1}{1-4\varepsilon^2}$$

$$= -\frac{1}{2}\log\left(1-4\varepsilon^2\right)$$

$$= -\frac{1}{2}\left(-4\varepsilon^2 + O(\varepsilon^4)\right)$$

$$= 2\varepsilon^2 + O(\varepsilon^4) = O(\varepsilon^2)$$

$$d_{KL}(P, Q) = O(m \varepsilon^2)$$

$$d_{TV}(P, Q) = O(\sqrt{m}\, \varepsilon)$$

Απόδειξη του Λήμματος του Le Cam

Error Prob =

$$\max \left\{ \Pr_{x \sim P}[T(x) = Q], \Pr_{x \sim Q}[T(x) = P] \right\}$$

$$\geq \frac{1}{2} \left(\Pr_{x \sim P}[T(x) = Q] + \Pr_{x \sim Q}[T(x) = P] \right)$$

$$= \frac{1}{2}\left(\sum_x P(x)\, 1[T(x)=Q] + \sum_x Q(x)\, 1[T(x)=P] \right)$$

$$= \frac{1}{2}\left(\sum_x P(x)\left(1 - 1[T(x)=P]\right) + \sum_x Q(x)\, 1[T(x)=P] \right)$$

$$= \frac{1}{2}\left(1 - \sum_x \left(P(x)-Q(x)\right) 1[T(x)=P] \right)$$

$$\geq \frac{1}{2}\left(1 - \sum_x \left(P(x)-Q(x)\right)_+ \right)$$

$$= \frac{1}{2}\left(1 - d_{TV}(P,Q) \right) \qquad \square$$