Вопросы для зачета по курсу
«Статистические модели распознавания» (май 2021)

1. Байесовские задачи распознавания. Детерминированный характер стратегий байесовского распознавания. Стратегии приближенного решения байесовских задач, основанные на минимизации эмпирического риска.

2. Неравенство Чебышева и его доказательство.
   Пусть μ -- математическое ожидание случайной величины x, σ^2 -- её дисперсия , $x_n^* = \frac{1}{n}\sum_{i=1}^{n} x_i$ -- среднее значение величины x_i на случайной выборке длины n. Вывести формулу для оценки сверху вероятности неравенства $|x_n^* - \mu| \geq \Delta$.

3. Пусть μ -- математическое ожидание случайной величины x, σ^2 -- её дисперсия , $x_n^* = \frac{1}{n}\sum_{i=1}^{n} x_i$ -- среднее значение величины x_i на случайной выборке длины n. На основании формулы для оценки сверху вероятности неравенства $|x_n^* - \mu| \geq \Delta$ вывести формулу для оценки сверху вероятности неравенства $|R^e(d) - R^t(d)| \geq \Delta$, где $R^t(d)$ -- риск решения d, $R^e(d)$ -- эмпирический риск заданного решения d на случайной выборке длины n.

4. Пусть $R^t(d)$ -- риск решения d, $R^e(d)$ -- эмпирический риск решения d на случайной выборке длины n, $d^e = \arg\min_d R^e(d)$. Вывести оценку сверху для величины $R^t(d^e) - \min_d R^t(d)$, если $\forall d \in D[|R^e(d) - R^t(d)| \leq \Delta]$.

5. Пусть $R^t(d)$ -- риск решения d, $R^e(d)$ -- эмпирический риск решения d на случайной выборке длины n, $d^e = \arg\min_d R^e(d)$. Вывести оценку сверху для вероятности того, что условие $\forall d \in D[|R^e(d) - R^t(d)| \leq \Delta]$ не выполнится, и на этом основании показать, как следует выбирать длину эмпирической выборки.

6. Сформулировать и обосновать необходимые и достаточные условия, которым должны удовлетворять условные вероятности $p(x|y)$ и $p(y|x)$, чтобы они однозначно определяли совместные вероятности $p(x,y)$.

7. Сформулировать «алгоритм» отыскания маргинальных вероятностей $p(x)$ и p(y) на основании известных условных вероятностей $p(x|y)$ и $p(y|x)$ и доказать его сходимость. Определить гиббсов самплер и сформулировать его основное свойство.

8. Экспоненциальные случайные велчины. Вывести общий вид распределения вероятностей экспоненциальной случайной величины.

9. Гиббсовы поля. Вывести общий вид распределения вероятностей гибсова поля.

10. Марковские свойства гиббсова поля и гиббсов самплер гиббсова поля.

11. Наиболее правдоподобное оценивание параметров экспоненциальной случайной величины.

12. Наиболее правдоподобное оценивание параметров гиббсова поля.

13. Сформулировать понятие рандомизированной стратегии и риска рандомизированной стратегии. Привести пример небайесовской задачи, для которой существует рандомизированная стратегия, риск которой лучше, чем риск любой детерминированной сиратегии.

14. Задачи распознавания при неполностью известной статистической модели распознаваемого обэъекта. Функция риска стратегии. Пригодные и непригодные стратегии.

15. Байесовы стратегии распознавания при статистической модели, известной с точностью до неизвестного параметра. Лемма о дихотомии байесовых и непригодных стратегий.

16. Привести примеры небайесовских задач распознавания: минимаксные задачи, (minimax deviation)-задачи и другие. Показать, что стратегии решения этих задач являются байесовскими, то есть, не являются непригодными.

17. Суперградиент вогнутой функции. Показать, чему равен суперградиент функций $a \cdot f$, $f_1 + f_2$, $\max\{f_1, f_2\}$, если известны суперградиенты функций f , f_1, f_2.

18. Сформулировать процесс субградиентного подъема и основное его свойство.

19. Сформулировать алгоритм отыскания стратегии решения минимаксных и (minimax deviation)-задач распознавания.

1.

X - ознака об'єкта, спостереження

$x \in X$ — мн-на спостережень

k - стан об'єкта (прихована вел-на)

$k \in K$ — мн-на станів

$p(x, k)$ — ймов. того, що об'єкт знаходиться в стані k, а ознака цього об'єкта притаманна x.

$\mathcal{D}$ — мн-на рішень, d - рішення

$W: K \times \mathcal{D} \to \mathbb{R}$ — штрафна ф-я. $W(k, d)$ — штраф, який виплачується, коли об'єкт знаходиться в стані k і прийнято рішення d.

$q: X \to \mathcal{D}$ — стратегія (кожному спостереженню видає рішення)

Байєсівська задача:

Для заданих $X, K, \mathcal{D}$, $p_{XK}: X \times K \to \mathbb{R}$, $W: K \times \mathcal{D} \to \mathbb{R}$, побудувати стратегію, яка мінімізує ризик:

$$R(q) = \sum_{x \in X} \sum_{k \in K} p_{XK}(x, k)\, W(k, q(x))$$

Рішенням байєсівської задачі є стратегія q^* — байєсівська стратегія

рандоміз. стратегія: $q_r: \mathcal{D} \times X \to \mathbb{R}$, ризик рандоміз. стратегії

$$R^{rand}(q) = \sum_{x \in X} \sum_{k \in K} p(x, k) \sum_{d \in \mathcal{D}} q_r(d|x)\, W(k, d).$$

входить лінійно

Теорема

$\exists$ детерм. стратегія $q: X \to \mathcal{D}$, $R^{det}(q) \leq R^{rand}(q_r)$

$$R^{rand} = \sum_{x \in X} \sum_{d \in \mathcal{D}} q_r(d|x) \sum_{k \in K} p(x, k)\, W(k, d)$$

$$\sum_{d \in \mathcal{D}} q_r(d|x) = 1 \quad \forall x \in X, \quad q_r(d|x) \geq 0 \quad \forall d \in \mathcal{D}, \forall x \in X$$

$$R^{rand} \geq \sum_{x \in X} \min_{d \in \mathcal{D}} \sum_{k \in K} p(x, k)\, W(k, d).$$

Розглянемо $q(x)^*$ значення d, для якого

$$\sum_{k} p(x,k)\, w(k, q(x)) = \min_{d\in D} \sum_{k\in K} p(x,k)\, w(k,d)$$

Визначена таким чином ф-я $q: X \to D$ є детермінована стратегія, яка не гірше, ніж рандомізована. $R^{rand} \geq \sum_{x\in X}\sum_{k\in K} p(x,k)\, w(k,q(x))$ ▲

Рандомізація не покращує рішення, але в деяких випадках буває корисною в силу своїх інших особливостей.

Розглянемо 2 випадки для яких рандомізація є логічним і обґрунтованим способом.

← точний ризик рішення d

$$x^* \longmapsto d^* = \arg\min_{d\in D} \sum_{k\in K} p(k|x^*) \cdot w(k,d)$$

$|D|\cdot|K|$ - складність

але в деяких випадках $|D|$ і $|K|$ є великими для безпосереднього обрахунку.

(I) Розглянемо випадок, коли $|K|$ - велике, $p(k|x^*)$ - обчислюється просто, $|D|$ - мале

Тоді генеруємо незалежним чином вибірку $k_1, k_2, \dots, k_n$ (у відповідності з розподілом $p(k|x^*)$

Тоді можна обчислити емпіричний ризик $\frac{1}{n}\sum_{i=1}^{n} w(k_i; d)$

$$\underbrace{\sum_{k\in K} p(k|x^*)\, w(k,d)}_{R^T(d)} \approx \underbrace{\frac{1}{n}\sum_{i=1}^{n} w(k_i; d)}_{R^e(d)}$$

Ідея:

Замість мінімізації точного ризику будемо мінімізувати емпіричний ризик.

1.

II) $x^{\circ} \longmapsto d^{\circ} = \underset{d \in D}{\arg\min} \sum_{k \in K} p(k/x^{\circ})\, \omega(k, d)$

d є функцією від k ($f(k)$ — приймає малу к-сть значень) $|D| << |K|$

$d = f(k)$, $p(k/x^*) \longmapsto \pi(d/x^*)$

нехай w задана як $w(d, d') = \omega(f(k), d')$

Якщо розподіл π задати, задача формулюється як

$$d^{\circ} = \underset{d'}{\arg\min} \sum_{d \in D} \pi(d/x^{\circ})\, \omega(d, d')$$

сума по малій кількості доданків.

Складність в тому, як побудувати $\pi(d/x^{\circ})$?

Прогенеруємо $k_1, \dots, k_n$ з розподілу $p(k/x^{\circ})$

Хочемо побудувати $d_1, \dots, d_n$ з розподілом $\pi(d/x^{\circ})$, але для цього $\pi(d/x^{\circ})$ навіть і не потрібно знати. Для цього достатньо $d_i = f(k_i)$. При такій побудові $d_1, \dots, d_n$ буде мати розподіл $\pi(d/x^{\circ})$.

На основі вибірки можемо відновити $\pi(d/x^{\circ})$ (частоти як оцінки)

2. Нерівність Чебишева і його доведення.
Нехай μ - мат. очікування випадкової величини X, σ^2 - її дисперсія, $X_n^* = \frac{1}{n}\sum_{i=1}^{n} x_i$ - середнє значення величини x_i на випадковій вибірці довжини n. Вивести формулу для оцінки зверху ймовірності нерівності $|X_n^* - \mu| \le \Delta$.

$$x \in X, \ p(x) \geq 0; \ \sum_{x \in X} p(x) = 1.$$

Нехай $f(x) \to \mathbb{R}$!

$$\mu = \sum_{x \in X} f(x) p(x) \text{ - мат. очікування } f(x).$$

$$\sigma^2 = \sum_{x \in X} (f(x) - \mu)^2 \cdot p(x) \text{ - дисперсія } f(x).$$

Зрозуміло, що $\sum_{x \in X} (f(x) - \mu) = 0$.

Якщо дисперсія мала, то ймовірність відхилення ф-ії $f(x)$ від μ теж мале

σ^2 задано.
Доведемо: $P(|f(x)| - \mu| \geq \Delta) \le \frac{\sigma^2}{\Delta^2}$ - нерівність Чебишева.

▲

$$\sum_{|f(x) - \mu| \geq \Delta} p(x) = \frac{1}{\Delta^2} \cdot \sum_{|f(x) - \mu| \geq \Delta} \Delta^2 p(x) \le \frac{1}{\Delta^2} \sum_{|f(x) - \mu| \geq \Delta} (f(x) - \mu)^2 \cdot p(x) \le$$

$$\le \frac{1}{\Delta^2} \sum_{x \in X} (f(x) - \Delta)^2 \cdot p(x) = \frac{\sigma^2}{\Delta^2}$$ ▲

Нехай $x \in X$, $p_x(x)$; μ_x і σ_x^2 і $y \in Y, p_y(y)$; μ_y і σ_y^2.

Є ф-я: $z = x + y$: $\mu_z = \mu_x + \mu_y$

$\sigma_z^2 = \sigma_x^2 + \sigma_y^2$, якщо x, y - незалежні.

Якщо $z = ax \Rightarrow z = a^2 \sigma_x$

Нехай $x \in X$, $p(x)$, μ_x, σ_x^2.

$(x_1, x_2, x_3, \ldots x_m)$ - вибірка взаємно нез-их величин

для неї визначено $\frac{1}{m}\sum_{i=1}^{m} x_i = x^m$; Обчислимо μ^m, σ^{2m}

$$\mu^m = \mu_x; \quad \sigma^{2m} = \frac{1}{m^2}\sum_{i=1}^{m}\sigma_m^2 = \frac{m\sigma_m^2}{m^2} = \frac{\sigma_m^2}{m}$$

Якщо $|x^m - \mu_x| \geq \Delta$: вважаємо, що вибірка „непридатна"

$$\Rightarrow P(|x^m - \mu| \geq \Delta) \leq \frac{\sigma^2}{m\Delta^2}$$ ← з нер-ті Чебишева.

3.

Вибірка $(k_1,\dots,k_m)$ називається <u>Δ-придатною</u>, якщо $\forall d \in D: |R^e(d) - R^T(d)| < \Delta$

-||- Δ-непридатною, якщо $\exists d \in D: |R^e(d) - R^T(d)| \geq \Delta$.

$x \in X$, $p(x), f(x) \geq 0$, $\sum_x p(x) = 1$,

$f: X \to \mathbb{R}$, $\mu = \sum_{x \in X} f(x) p(x)$; $\sigma^2 = \sum_{x \in X} (f(x) - \mu)^2 p(x)$

Нерівність Чебишева: $P(|f(x) - \mu| \geq \Delta) \leq \frac{\sigma^2}{\Delta^2}$

$x \in X$, $p(x)$, μ_x, σ_x^2

$(x_1,\dots,x_m) \longmapsto \frac{1}{m} \sum_{i=1}^{m} x_i = X^m$, μ^m, σ^{2m}

$\mu^m = \mu_x$, $\sigma^{2m} = \frac{\sigma_x^2}{m}$ $\quad |X^m - \mu_x| \geq \Delta$ – непридатна вибірка.

$$\boxed{P(|X^m - \mu_x| \geq \Delta) \leq \frac{\sigma_x^2}{m \cdot \Delta^2}}$$

Використаємо в контексті наших понять (ризик, емп. ризик і т.д.)

. $\omega(k,d)$ – втрати, які отримує система, коли приймає рішення d відносно об'єкта, який знаходиться в стані k.

При фіксованому d $w(k,d)$ – вип. величина. (т.як k – в.в.)

$\Rightarrow$ $\omega(k,d)$ має мат. очікування і дисперсію $= \sigma_w^2(d)$

Дисперсія невідома, але можемо ввести якусь зверху для дисперсії

$\exists\, \sigma_w^2 : \forall d \, [\, \sigma_w^2(d) - \sigma_w^2 \leq 0 \,]$

наприклад: $\sigma_w^2 = \max w(k,d) - \min w(k,d)$

$$R^T(d) = \sum_{k\in K} p(k|x^0)\cdot w(k,d)$$

генеруємо $k_1,\dots,k_m$ з $p(k|x^0)$ розподілу.

$$(k_1,\dots,k_m) \longmapsto R^e(d) = \frac{1}{m}\sum_{i=1}^{m} w(k_i,d)$$

$$\boxed{P\left(|R^T(d)-R^e(d)| \geq \Delta\right) \leq \frac{\sigma_w^2}{m\cdot\Delta^2}}$$

4. Нехай $R^t(d)$ - ризик рішення d, $R^e(d)$ - емпіричний ризик рішення d на випадковій вибірці довжини n, $d^e = \arg\min_d R^e(d)$. Вивести оцінку зверху для величини $R^t(d^e) - \min_d R^t(d)$, якщо $\forall d \in D\,[\,|R^e(d) - R^t(d)| \le \Delta\,]$

$x \in X$, $k \in K$; (x,k) - випадкова пара.
Для кожного значення x та k відомо $p(x,k)$, також відома апостеріорна ймовірність $p(k/x)$
$d \in D$ - кінцева мн-на рішень системи розпізнавання

$q: X \to D$; $q(x) \in D$ стратегія рішень (може бути хорошою чи поганою)

$w: K \times D \to \mathbb{R}$; $w(k,d)$ - штраф, що виплачує система розпізнавання.

$$R^T(d) = \sum_{k \in K} p(k/x)\, w(k,d)$$ ← ризик (мат. очікування) рішення d

$R^T: D \to \mathbb{R}$ - точна ф-я ризика

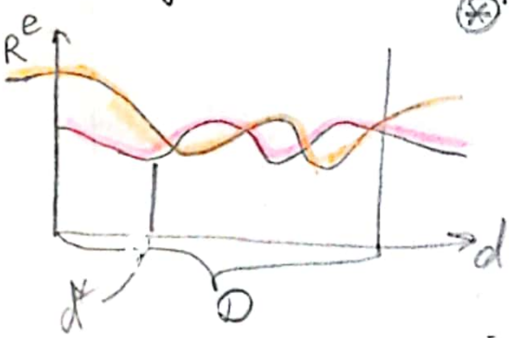
Задача: $d^* = \arg\min_d R^T(d)$ (выбрать те рішення, що мінімізує ризик.

Хочемо замінити мін-ю R^T на мін-ю R^e

Нехай програма видає вибірку $(k_1, \dots, k_m)$

$$\underbrace{R^e(d) = \frac{1}{m}\sum_{i=1}^{m} w(k_i, d)}$$ - ф-я емпіричного ризику
← середнє арифм.

З ростом m R^e стає не відрізнюваним від R^T в тому сенсі, що (*) $|R^e(d) - R^T(d)| < \Delta_0$ (для великих m).



(Ймовірність великих відхилень даних величин стає малою).

Def
Вибірку $(k_1, \dots, k_n)$ назвемо Δ-придатна, якщо
$\forall d \in D$: $|R^e(d) - R^T(d)| < \Delta$

Def
Вибірка $(k_1, \dots, k_n)$ наз. не придатною, якщо $\exists d \in D$:
$|R^e(d) - R^T(d)| \geq \Delta$

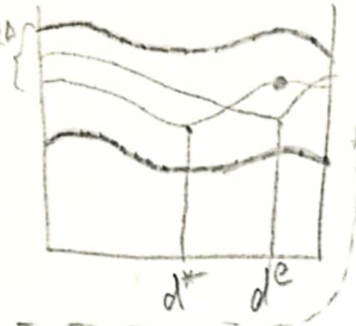
якщо R^e лежить в коридорі => вибірка придатна.

Для повного розуміння використання R^e замість R^T потрібно розглянути 2 питання:

1) Якщо вибірка Δ-придатна, то що потрібно поставити замість „?": $|R^T(d^e) - R^T(d^*)| <$ „?"

2) Для заданого m (об'єму вибірки): Яка ймовірність не придатних вибірок, т.$\mathcal{E}$ – сумарна ймов-ть всіх вибірок, що не придатні. Потрібно вказати яким чином $\mathcal{E}$ залежить від m і інших параметрів: $\mathcal{E} < Q(m, \dots)$

Розглянемо 1) (по білету):
Вибірка придатна; На скільки відрізняється $R^T(d^*)$ і $R^T(d^e)$?

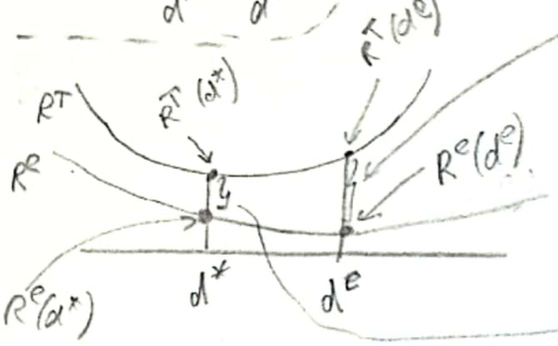


Насправді $|R^T(d^e) - R^T(d^*)| < 2\Delta$

Доведемо: ▲

Потрібно розглянути 3 положення R^e і R^T

1) Обидві точки $R^e(d^*)$ і $R^e(d^e)$ нижчі за $R^T(d^*)$ і $R^T(d^e)$

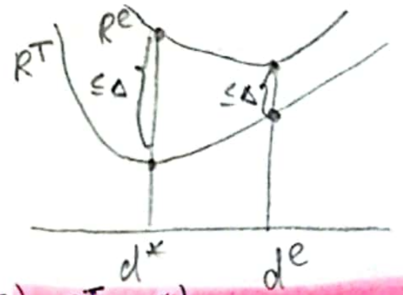


це різниця не перевищує Δ
тобто $|R^e(d^e) - R^T(d^e)| < \Delta$ по (*)

$R^e(d^e)$ нижче за $R^e(d^*)$ =>
$R^e(d^*)$ вище за $R^e(d^e)$
=> $|R^T(d^*) - R^e(d^*)| < |R^T(d^e) - R^e(d^e)|$

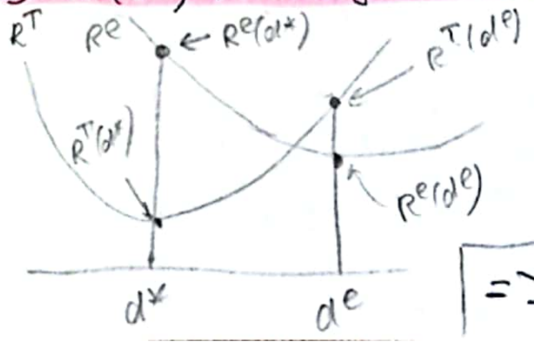
=> $|R^T(d^*) - R^T(d^e)| \le \Delta$

2) Обидві точки $R^e(d^*)$ і $R^e(d^e)$ вищі за $R^T(d^*)$ і $R^T(d^e)$



→ Аналогічно до 1)
=> $|R^T(d^*) - R^T(d^e)| \le \Delta$

3) $R^T(d^*)$ нижче за $R^e(d^e)$; $R^e(d^*)$ вище за $R^T(d^*)$ і $R^e(d^e)$



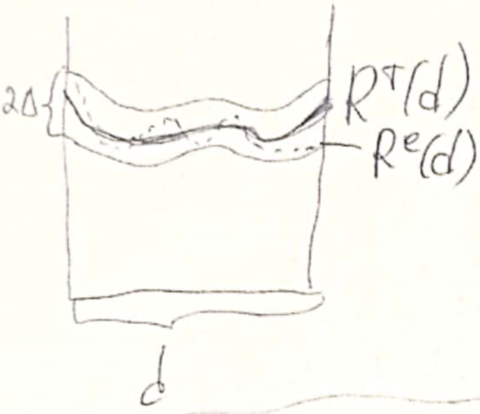
$R^e(d^*)$ не може бути нижче за $R^e(d^e)$,
а $|R^e(d^*) - R^T(d^*)| < \Delta$ і =>
$|R^T(d^e) - R^e(d^e)| < \Delta$
=> $|R^T(d^*) - R^T(d^e)| \le 2\Delta$ ▲

$\boxed{\Rightarrow |R^T(d^*) - R^T(d^e)| \le 2\Delta}$ якщо вибірка придатна!

№ 5.

$R^T(d) = \sum_{k\in K} p(k|x^0)\cdot W(k,d)$ $\quad R^e(d) = \frac{1}{m}\sum_{i=1}^{m} W(k_i; d)$

Δ-прийн. вибірка: $\forall d\in D\,[\,|R^T(d) - R^e(d)| < \Delta\,]$ (для кожної d)



Яка ймовірність Δ-неприйнятної вибірки?

$P(\exists d: [\,|R^T(d) - R^e(d)| \geq \Delta\,]) - ?$

розглядаємо всю функцію $R^T(d), R^e(d)$.

Відомо, що $P(|R^T(d) - R^e(d)| \geq \Delta) \leq \frac{\sigma_w^2}{m\cdot\Delta^2}$

мн-на всіх вибірок

мн-на вибірок, які є „поганими" для d_1

сумарна ймовірність $\leq \frac{\sigma_w^2}{m\cdot\Delta^2}$

тоді те, що ми шукаємо це мн-на  (всі „погані" вибірки при якихось d)

$$P(\exists d: [\,|R^T(d) - R^e(d)| \geq \Delta\,]) \leq \sum_{d\in D} P(|R^T(d) - R^e(d)| \geq \Delta) \leq$$

$$\leq \sum_{d\in D} \frac{\sigma_w^2}{m\cdot\Delta^2} = \frac{|D|\cdot\sigma_w^2}{m\,\Delta^2}$$

Scanned with CamScanner

$$\boxed{P[\exists d[|R^T(d) - R^e(d)| \geq \Delta]) \leq \frac{|D| \cdot \sigma_w^2}{m \cdot \Delta^2}}$$

$|D|$ - к-сть можливих рішень
σ_w^2 - оцінка зверху для дисперсії
m - довжина вибірки
Δ - допустима похибка.

В: Як можна використати це на практиці?

$\frac{\sigma_w}{\Delta}$ - відносна похибка. Нехай 1% - ~~похибка~~ знайдемо точний ризик оцінимо з помилкою $\leq \Delta$.

$$\frac{\sigma_w}{\Delta} = 10^2$$

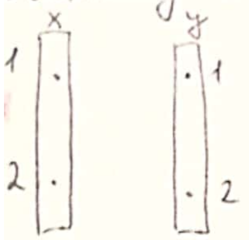
Також хочемо, щоб ймовірність неприйнятної вибірки була малою. Нехай також ймовірність виходу за межі: $P(\exists d\, |R^T(d) - R^e(d)| \geq \Delta) \leq \frac{|D| \cdot \sigma_w^2}{m \cdot \Delta^2} \leq 10^{-2}$

$$\frac{|D| \cdot \sigma_w^2}{m \cdot \Delta^2} \leq 10^{-2} \qquad \frac{|D|}{m} \cdot 10^6 \leq 1, \quad \boxed{m \geq |D| \cdot 10^6}$$

При великих $|D|$ ~~m~~ m може бути також дуже великим, але в деяких випадках це явно краще, ніж мінімізація точного ризику.

6. Сформулювати і обґрунтувати необхідні і достатні умови, яким мають задовольняти ум. ймовірності $p(x|y)$ і $p(y|x)$, щоб вони однозначно визначали спільні ймов-ті $p(x,y)$

Нехай у нас $x \in X$, $y \in Y$ — випадкові величини.

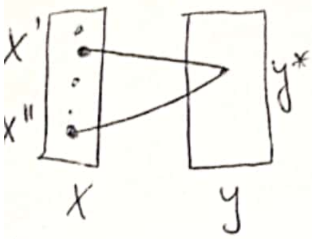


$p(x,y)$ - визначає ймовірність пари x, y.

$p(x,y) = p(x) \cdot p(y|x) = p(y) \cdot p(x|y)$.

У загальному випадку $p(y|x)$, $p(x|y)$ не визначає однозначно $p(x,y)$, $p(x)$, $p(y)$

Нам потрібно описати випадки, коли вони точно визначають



Визначимо граф Γ_x

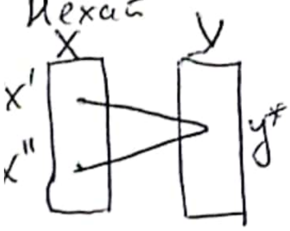
x', x'' - сусіди, якщо $\exists y^* \in X [p(x',y^*) \neq 0, p(x'',y^*) \neq 0]$

Теорема

Якщо Γ_x складається із єдиної зв'язної компоненти, то $p(y|x)$ і $p(x|y)$ точно визначають $p(x,y)$, а відповідно і $p(x), p(y)$.

А якщо Γ_x складається із багатьох зв'язних компонент, то $p(y|x)$ і $p(x|y)$ не визначають $p(x,y)$ і відповідно $p(x), p(y)$.

▲ Нехай



Існує пара x', x'' сусідів, що $p(x',y^*) \neq 0$, $p(x'',y^*) \neq 0$

$$\frac{p(x'/y^*)}{p(x''/y^*)} = \begin{cases} p(x',y^*) = p(x') \cdot p(y^*/x') = p(y^*) \cdot p(x'/y^*) \\ p(x'',y^*) = p(x'') \cdot p(y^*/x'') = p(y^*) \cdot p(x''/y^*) \end{cases}$$

$$= \frac{p(x') \cdot p(y^*/x) / \cancel{p(y^*)}}{p(x'') \cdot p(y^*/x'') / \cancel{p(y^*)}} = \frac{p(x')}{p(x'')} \cdot \frac{p(y^*/x')}{p(y^*/x'')}$$

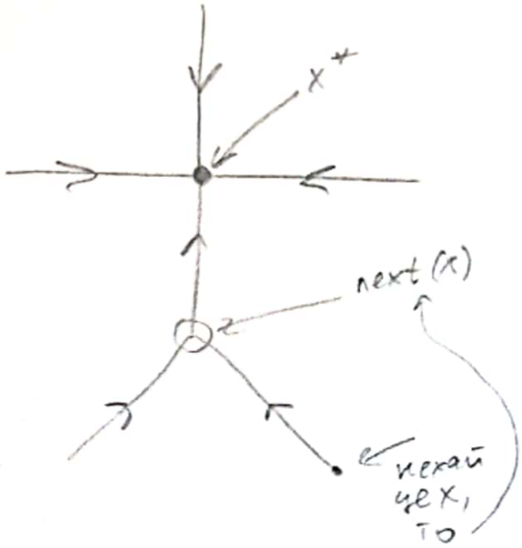
$\Rightarrow$ Для двух сусідів: $\dfrac{p(x')}{p(x'')} = \dfrac{p(x'/y^*) \cdot p(y^*/x'')}{p(x''/y^*) \cdot p(y^*/x')}$ (*)

Напишемо алгоритм для відновлення $p(X)$ на основі $p(X|Y)$ та $p(Y|X)$.

$\Rightarrow$ Якщо граф Γ_X складається із єдиної зв'язної компоненти то в ньому існує зв'язне дерево.

1) Побудувати зв'язне дерево

2) В ньому обрати будь-яку вершину x^* як корінь.



Зробити це дерево орієнтованим (визначити стрілки до кореня)

$next(x)$ - вершина по стрілці яка веде від x в напрямку кореня

3) Присвоїмо кореню $g(x^*) = 1 = \underset{\text{якесь число}}{c} \cdot p(x^*)$

4) $x \in X$ - це така вершина для якої $g(next(x)) = c\,p(next(x))$

ймовірності не відомі але відоме відношення (*)

$$\Rightarrow g(x) = g(next(x)) \cdot \frac{p(x)}{p(next(x))} = c \cdot p(x)$$

Знаючи числа g ми можемо вказати $p(x) = \dfrac{g(x)}{\sum_x g(x)}$.

$\Leftarrow$ без дов.

В такому випадку, якщо багато зв'язних компонент то пара ф-ій $p(x|y), p(y|x)$ не визначають точно $p(x,y)$
(„а визначає з точністю якогось <u>параметру</u>: по суті це буде сума зв'язних компонент")

7. П. 1

$$[P(x/y), p(y/x)] \longmapsto (p(x), p(y))$$

Те, що ми шукаємо в рішеннях лінійної сист. рівнянь

$$
\begin{array}{ll}
p_x^o(x) & \begin{cases} p(x) = \sum\limits_{y \in Y} p(y)\, p(x/y), & x \in X \quad (1) \\ p(y) = \sum\limits_{x \in X} p(x)\, p(y/x), & y \in Y \quad (2) \end{cases} \\
p_y^o(y) &
\end{array}
$$

Алгоритм не можна застосовувати на практиці, т.як він вимагає всі значення $x \in X, y \in Y$ - а це може бути неможливо.

Система рівнянь вирішується наступним чином:

- беремо довільний розпр. $p(x)$
- підставляємо $p(x)$ в (2), отримуємо $p^1(y)$
- підставляємо $p^1(y)$ в (1), отримуємо $p^2(x)$

"ітерація алгоритму"

$$p_0(x) \to p_1(y) \to p_1(x) \to \dots \; p_i(y) \to p_i(x)$$

Нехай $p_y^o(y), p_x^o(x)$ - розв'язок системи р-нь.

Тоді: $$\lim_{i\to\infty} \max_{x \in X} |p_i(x) - p_x^o(x)| = 0.$$

Алгоритм також є незастосовним на практиці.

Але можемо використати підхід для отримання алгоритму, який не вимагає такого обсягу пам'яті.

- візьмемо довільне x_0
- генеруємо $y_1 = \text{rand}\, p(y/x_0)$ } ітерації
- генеруємо $x_1 = \text{rand}\, p(x/y_1)$ }

. . .

y_1 з розподілу $p_1(y)$, ...
На кожному кроці отримуємо ~~вже~~ не набір ймовірностей, а випадкову величину із заданого розподілу.
В результаті достатньо довгого такого процесу, ми отримаємо $x_i \mapsto \text{rand}\, p_x^*(x)$.

Збіжність алгоритму (доведення)

$x \in X,\ y \in Y,\quad p(x,y) \quad p_x(x/y),\ p_y(y/x)$

$x^0 \to p_{y/x}(y/x^0) : y^1$

$y^1 \to p_{x/y}(x/y^1) : x^1 \quad \dots \longrightarrow x^k \to y^{k+1} \to x^{k+1}$

у x^k розподіл ймов. $p_x^k(x)$. тоді

$$p_y^{k+1}(y) = \sum_{x' \in X} \underbrace{p_{y/x}(y/x')\, p_x^k(x')}_{p_{x,y}(x',y^*)} \qquad (1)$$

$$p_x^{k+1}(x) = \sum_{y' \in Y} p_{x/y}(x/y') \cdot p_y^{k+1}(y') \qquad (2)$$

Хочемо отримати залежність $p^{k+1}(x)$ $(p^k(x))$. Підстав. 1 в 2.

$$p_x^{k+1}(x) = \sum_{y' \in Y} p_{x/y}(x/y') \cdot p_y^{k+1}(y') = \sum_{y' \in Y} p_{x/y}(x/y') \cdot \sum_{x' \in X} p_{y/x}(y'/x') \cdot p_x^k(x') =$$

$$\cancel{= \sum_{x' \in X} \sum_{y' \in Y} p_{x/y}(x/y') \cdot p_{y/x}}$$

7. П.2

$$\ominus \sum_{x'\in X}\underbrace{\sum_{y'\in Y} P_{x|y}(x/y')\cdot P_{y|x}(y'/x')}_{A_{xx'}}\cdot p_x^k(x') = p_x^{k+1}(x) \ominus$$

зв'язує наступн. від попер

$|X|=n$

$$\ominus \sum_{x'\in X} A_{xx'}\, p_x^k(x') \leftarrow \text{мн-ня матриці на вектор.}$$

X – скінченна мн-на $\to$ перенумеруємо їх від 1 до $|X|$.

Тоді $p_x^k(1) = p_1^k$, $p_x^k(2) = p_2^k$ $\quad \bar p^k$, $\quad \boxed{\bar p^{k+1} = A\bar p^k}$

Покажемо, що при деяких умовах для A процес сходиться.

Потрібні вл-ті для A.

1. $\min_{i,j} A_{ij} = \alpha > 0$ (всі ел-ти A – ймовірності ≥ 0)

2. $\forall \bar z \in \mathbb{R}^n: \sum_{i=1}^n z_i = \sum_{i=1}^n (Az)_i$

▲

$$A = \begin{array}{c} x' \\ \begin{pmatrix} - & - & - & - \\ \cdot & - & - & - \\ \cdot & - & - & - \end{pmatrix} \\ \| \\ 1 \end{array}$$

$$\sum_{x\in X}\sum_{y'\in Y} P_{x|y}(x/y')P_{y|x}(y'/x') = \sum_{y'\in Y}\Big[\underbrace{\sum_{x\in X} P_{x|y}(x/y')}_{=1}\Big]P_{y|x}(y'/x') = \underbrace{\sum_{y'\in Y} P_{y|x}(y'/x')}_{=1}$$

$$\sum_{i=1}^n (A\bar z)_i = \sum_{i=1}^n\sum_{j=1}^n A_{ij} z_j = \sum_{j=1}^n z_j \underbrace{\sum_{i=1}^n A_{ij}}_{=1} = \sum_{j=1}^n z_j \quad ▲$$

3. $\forall \bar z: \bar z \geq 0: (A\bar z)_i \geq 0, \ \forall i=\overline{1,n}$

▲ $A_{ij} \geq 0 \ \forall i,j \Rightarrow (A\bar z)_i \geq 0 \ \forall i=\overline{1,n}$ ▲

Хочемо показати, що $\bar{\Delta}^k = \underbrace{A\bar{p}^k}_{\bar{p}^{k+1}} - \bar{p}^k \to 0$

4. $\sum_{i=1}^{n} \Delta_i^k = 0 \quad \forall k \in N^+$ за власт. (2) $= \sum_{i=1}^{n} p_i^k$

▲ $\sum_{i=1}^{n} \Delta_i^k = \sum_{i=1}^{n} (A\bar{p}^k)_i - \sum_{i=1}^{n} p_i^k = 0$ ▲

$\bar{\Delta} = \bar{\Delta}^+ - \bar{\Delta}^-, \quad \Delta_i^+ = \begin{cases} \Delta_i, & \Delta_i \geq 0 \\ 0, & \Delta_i < 0 \end{cases}, \quad \Delta_i^- = \begin{cases} -\Delta_i, & \Delta_i < 0 \\ 0, & \Delta_i \geq 0 \end{cases}$

5. $\sum_{i=1}^{n} \Delta^+ = \sum_{i=1}^{n} \Delta_i^- = \frac{1}{2} \sum_{i=1}^{n} |\Delta_i|$

▲ очевидно ▲

6. $a \geq 0, \; b \geq 0, \; |a-b| = a+b-2\min\{a,b\}$

▲ $a > b$: $|a-b| = a+b-2\min\{a,b\} = a+b-2b = a-b$

$a \leq b$: $|a-b| = a+b-2\min\{a,b\} = a+b-2a = b-a$ ▲

$\bar{\Delta}^k = A\bar{x}^k - \bar{x}^k$

$\bar{\Delta}^{k+1} = A\bar{x}^{k+1} - \bar{x}^{k+1} = AA\bar{x}^k - A\bar{x}^k = A(A\bar{x}^k - \bar{x}^k) = A\bar{\Delta}^k$.

Тоді розглянемо величину:

$$\sum_{i=1}^{n} |(A\bar{\Delta})_i| = \sum_{i=1}^{n} |(A\bar{\Delta}^+ - A\bar{\Delta}^-)_i| = \sum_{i=1}^{n} |(A\bar{\Delta}^+)_i - (A\bar{\Delta}^-)_i| =$$

$$= \sum_{i=1}^{n} \left((A\bar{\Delta}^+)_i + (A\bar{\Delta}^-)_i - 2\min\left\{ (A\bar{\Delta}^+)_i, (A\bar{\Delta}^-)_i \right\} \right) =$$

(власт. (1))

$$= \sum_{i=1}^{n} (A\bar{\Delta}^+)_i + \sum_{i=1}^{n} (A\bar{\Delta}^-)_i - 2\sum_{i=1}^{n} \min\left\{ \sum_{j=1}^{n} \overset{\geq \alpha}{A_{ij}} \Delta_j^+, \sum_{j=1}^{n} \overset{\geq \alpha}{A_{ij}} \Delta_j^- \right\} \leq$$

$$\leq \sum_{i=1}^{n} \Delta_i^+ + \sum_{i=1}^{n} \Delta_i^- - 2\sum_{i=1}^{n} \min\left\{ \sum_{j=1}^{n} \alpha \Delta_j^+, \sum_{j=1}^{n} \alpha \Delta_j^- \right\} =$$

$$= \sum_{i=1}^{n} |\Delta_i| - 2\sum_{i=1}^{n} \min\left\{ \frac{1}{2}\alpha \sum_{j=1}^{n} |\Delta_j|, \frac{1}{2}\alpha \sum_{j=1}^{n} |\Delta_j| \right\} =$$

7.

$$= \sum_{i=1}^{n}|\Delta_i| - \underbrace{\sum_{i=1}^{n}\alpha\sum_{j=1}^{n}|\Delta_j|} = \sum_{i=1}^{n}|\Delta_i| - n\cdot\alpha\cdot\sum_{i=1}^{n}|\Delta_i| =$$

$$= \sum_{i=1}^{n}|\Delta_i|(1-n\alpha)$$

Сума модулів зменшилася в $(1-n\alpha)$ разів.

Відповідно $\bar{\Delta}^k \to \bar{0}$.

$\bar{p}^k \to \bar{p}^*$ нерухома точка $\quad A\bar{p}^k = \bar{p}^*$.

$$\bar{p}^k = \underbrace{AAA\ldots A}_{k \text{ разів}}\bar{p}^0 \qquad A_{xx'} = p_{x/x'}(x/x')$$

8. Експоненційні випадкові величини. Вивести загальний вигляд розподілу ймовірностей екс. вип. величини

Дано: X ; $p(X)$ невідомо ; $X \in \{0,1,\dots,9\}$;

$x \in X, y \in Y$, $p(x)$; $p(y)$; $p(x,y) = p(x) \cdot p(y)$.

$x \in \mathbb{R}$; $\int_{-\infty}^{+\infty} x p(x)\, dx = 0$, $\int_{-\infty}^{+\infty} x^2 p(x)\, dx = 1$ мат. очікування = 0, дисперсія = 1

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} = N(0,1).$$

Така гаусівська величина з (0,1) має найбільшу ентропію (тобто можна передати максимальний об'єм інформації через $N(0,1)$)

$x \in X$, $p(x)$; $f_1, f_2, \dots, f_n$! $f_i: X \to \mathbb{R}$ базис експ. вип. величини

Ці фії f_i задані, відомо їх мат очікування і дисперсії.

Напишемо систему p:

$$p: \begin{cases} \sum_{x \in X} f_i(x) \cdot p(x) = \Theta_i, \quad i = 1, \dots n \\ \sum_{x \in X} p(x) = 1 \end{cases}$$

параметри експ. величини

ентропія

$$p^* = \arg\max \left[- \sum_{x \in X} p(x) \cdot \log p(x) \right]$$

Експоненційна випадкова величина в базисі f з параметрами Θ наз. вип. величиною з максимальною ентропією і задов. цій системі.

Чи визначає воно точно? Так, бо це мн-на випукла а ентропія строго увігнута.

Максимум ентропії знаходиться за допомогою множників Лагранжа.

$$\Phi(p, \lambda_1, \dots, \lambda_n, \lambda_0) = - \sum_{x \in X} p(x) \log p(x) +$$

$$+ \sum_{i=1}^{n} \lambda_i \left(\sum_{x \in X} f_i(x) \cdot p(x) - \Theta_i \right) + \lambda_0 \left(\sum_{x \in X} p(x) - 1 \right)$$

Розглянемо деяке x^*

$$\frac{\partial \Phi}{\partial p(x^*)} = -\log p(x^*) - 1 + \sum_{i=1}^{n} \lambda_i \cdot f_i(x^*) + \lambda_0 = 0$$

$$\sum_{i=1}^{n} \lambda_i f_i(x^*) + \lambda_0 - 1 = \log p(x^*)$$

$$e^{\sum_{i=1}^{n} \lambda_i f_i(x^*) + \lambda_0 - 1} = p(x^*).$$

$e^{\lambda_0 - 1} = C$; замінимо $\lambda = (\lambda_1, \dots \lambda_n)$

$f = (f_1, \dots, f_n)$

$$\Rightarrow p(x^*) = C \cdot \exp[\lambda \cdot f(x)]$$

вектора.

9. П.1

T - множина об'єктів,

K - множина розміток $k(t)$ - мітка в об'єкті t, $t \in T$

$k: T \to K$, $k \in K^T$,

$T' \subset T$, T' - віконце. $k(T'): T' \to K$ (розмітка віконця)

$$k(T')(t) = \begin{cases} k(t), & t \in T' \\ \text{не визначено}, & \text{інакше} \end{cases}$$

$\tau = \{T_1', T_2', \dots, T_m'\}$ - підмножина віконець

τ - структура мн-ни T. $\tau \subset 2^T$

$\max\limits_{T' \in \tau} |T'|$ - порядок структури τ

Часто розглядаємо структури 2-го порядку.

$T' = \{t, t'\}$ $\{t, t'\} \subset T$, $\{t, t'\} \in \tau$ (тоді t і t' називають сусідами) $(tt' \in \tau)$

$N(t) = \{t' \in T \mid tt' \in \tau\}$

окіл об'єкта t

k - вип. в-на, $p(k)$ - ? $T, \tau, p_{T'}(k(T'))$ ← локальні ймовірності, $T' \in \tau$, $k(T') \in K^{T'}$

Як повязані локальні ймовірності з $p(k)$, яке невідоме? (відоме)

$$P \begin{cases} p_{T'}(k') = \sum\limits_{\substack{k \in K^T \\ k(T') = k'}} p(k), \quad T' \in \tau, \ k' \in K^{T'} & (\lambda(T', k')) \\ 1 = \sum\limits_{k \in K^T} p(k) & (\lambda_0) \end{cases}$$

Але існує багато $p(k)$, які задовольняють систему. Щоб вибрати одне, виберемо ту, яка має max ентропію.

$$p^0 = \arg\max_{p\in P}\Big[-\sum_{k\in K^T} p(k)\log p(k)\Big]$$

→ Шукане випадкове поле на мн-ні об'єктів T із заданою структурою τ і заданими локальними ймовірностями.

Якщо система має рішення, то P - опукла множина, ентропія - увігнута $\Rightarrow p^0$ - єдине

Розв'яжемо з доп. ф-ії Лагранжа.

$$\Phi(p,\lambda) = -\sum_{k\in K^T} p(k)\log p(k) + \underbrace{\sum_{T'\in\tau}\underbrace{\sum_{k'\in K^{T'}}\lambda(T',k')\underbrace{\sum_{\substack{k\in K^T\\ k(T')=k'}} p(k)}_{F_2}}_{F_1}}_{F_0} + \lambda_0\sum_{k\in K^T} p(k)$$

$$\boxed{\forall k^0\ \frac{\partial \Phi}{\partial p(k^0)} = 0.}$$

Розглянемо д-ті доданки окремо (F_0).

$$\frac{\partial F_2}{\partial p(k^0)} = \begin{cases} 1, & k^0(T') = k' \\ 0, & k^0(T') \neq k' \end{cases}$$

(k', T' - фіксовані)

$$\frac{\partial F_1}{\partial p(k^0)} = \lambda(T', k^0(T'))$$

$$\frac{\partial F_0}{\partial p(k^0)} = \sum_{T'\in\tau}\lambda(T', k^0(T'))$$

Тепер розглянемо весь вираз:

Scanned with CamScanner

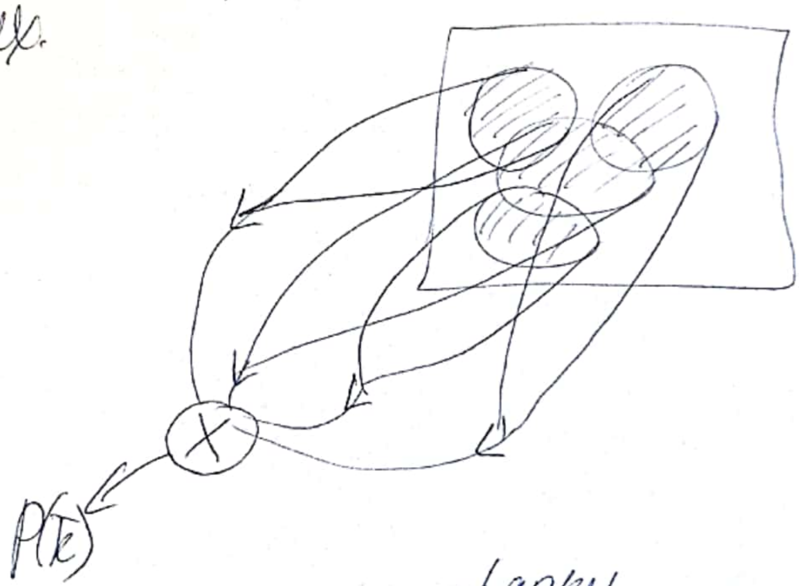
9. П. 2

$$\frac{\partial \Phi(p,\lambda)}{\partial p(k^0)} = -\log p(k^0) - 1 + \sum_{T'\in\tau} \lambda(T', k^0(T')) + \lambda_0 = 0 \quad \forall k^0 \in K$$

$$\log p(k) = C + \sum_{T'\in\tau} \lambda(T', k(T'))$$

$$\boxed{p(k) = \prod_{T'\in\tau} g_{T'}(k(T'))}$$

Ф-я великої к-сті змінних виражається як добуток ф-ій малої к-сті змінних.



Надалі розглядатимо структури 2-го порядку.

$$\boxed{p(k) = \prod_{tt'\in\tau} g_{tt'}(k_t, k_{t'})}$$

10. Марківські властивості гіббсового поля і гіббсовий самплер гіббсового поля

T - множина об'єктів, K - множина міток; $\bar{k}: T \to K$ - розмітка

$\bar{k} \in K^T$ - множина всіх розміток; $k(t)$ - мітка об'єкта t, $t \in T$

$T' \subset T$ - віконце

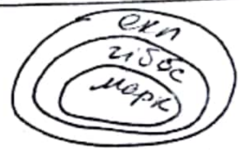
$k(T')$ - фрагмент розмітки k на віконці T' (звуження ф-ії $\bar{k}$ на підмножину T')

$k(T'): T' \to K$

$k(T')(t) = \begin{cases} k(t), & t \in T' \\ \text{не визначено для} & t \notin T' \end{cases}$

Гібсове випадкове поле часто ототожнюють з марківськими, але це не зовсім так і між ними є різниця.

Насправді Марківські вип. поле є підмножиною гіббсових випадкових полів.



Доводемо це для структур 2-го порядку

тобто коли кожне віконце складається тільки із двох об'єктів і ці об'єкти сусіди

Нехай $\mathcal{T} \subset T \times T$, $tt' \in \mathcal{T}$, $N(t^*) = \{t \mid tt^* \in \mathcal{T}\}$ - мн-на сусідів t^*

$\bar{N}(t^*) = \{t \neq t^* \mid tt^* \notin \mathcal{T}\}$ - мн-на не сусідніх об'єктів з t^*

$\mathcal{T}$ - множина всіх пар сусідніх об'єктів.

У гіббсовому випадковому полі нас цікавить

$$p(\bar{k}) = \prod_{tt' \in \mathcal{T}} g_{tt'}(k(t), k(t'))$$

складна ф-я великої кількості змінних.

У марківському випадковому полі нас цікавить

$$p(k(t^*) / k(T \setminus \{t^*\})) = p(k(t^*) / k(N(t^*)), k(\bar{N}(t^*))) \ominus$$

$k(T \setminus \{t^*\})$ - розмітка не сусідів t^*

$k(N(t^*))$ - розмітка сусідів t^*

$k(\bar{N}(t^*))$ - розмітка не сусідів t^*

Марковість насправді означає, що мітка об'єкта t^* насправді не залежить від $k(\bar{N}(t^*))$

$p(k(t^*)) = p(k(t^*)/k(\bar{N}(t^*)))$ – не правильне твердження.

"Якщо задикувати (прибити гвоздками) t^* і сусідів t^* то t^* залежить тільки від сусідів і не залежить від не сусідів".

$$\ominus \left\{ \begin{matrix} \text{по} \\ \text{Байесу} \end{matrix} \right\} = \frac{p(k(t^*), k(\bar{N}(t^*)), k(\bar{N}(t^*)))}{\sum_{k(t^*) \in K} p(k(t^*), k(N(t^*)), k(\bar{N}(t^*)))} =$$

$= \left\{ \begin{matrix} \text{Виділимо множину } T^N = \{t\,t^* \in T\} \text{ – множина пар з сусідами } t^* \\ \text{і } T^{\bar{N}} = T \setminus T^N \text{ – множина всіх решти пар} \end{matrix} \right\}$

$$= \frac{\prod_{t \in N(t^*)} g_{tt^*}(k(t), k(t^*)) \cdot \prod_{tt' \in T^{\bar{N}}} g_{tt'}(k(t), k(t'))}{\sum_{k(t^*) \in K} \prod_{t \in N(t^*)} g_{tt^*}(k(t), k(t^*)) \cdot \prod_{tt' \in T^{\bar{N}}} g_{tt'}(k(t), k(t'))} =$$

це не залежить від $k(t^*)$ то можна винести і скоротити

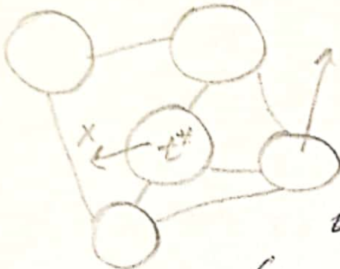
$$= p(k(t^*)/k(N(t^*)))$$

=> ми довели
$$p(k(t^*)/k(T \setminus \{t^*\})) = C \cdot \prod_{tt' \in T} g_{tt^*}(k(t), k(t^*))$$

=> марк. вип. поле це підмножина гібсового вип. поля. ▲

Гіббсовий семплер гіббсового поля

T - множина об'єктів; K - множина міток; $\bar{k}: T \to K$ - розмітка.
$k(t)$; $\tau \subset T \times T$ - структура сусідства;
$N(t) = \{t' \in T \mid tt' \in \tau\}$
Виходимо з того, що ф-я $\bar{k}$ (розмітка) є вип. гіббсовим полем:
=> Існує $p(\bar{k}) = \prod_{tt' \in \tau} g_{tt'}(k(t), k(t'))$



← Кожен з цих об'єктів знаходиться в якомусь стані, але ці стани приховані.
Для кожного об'єкта заданий сигнал, що він випромінює $x(t)$; $\bar{x}: T \to K$.
Сигнал залежить від стану об'єкту, причому якщо стан об'єкта заданий, то він не залежить від станів всіх інших об'єктів.
Це залежність x від k задається $p_t(x/k)$ ← відомо.

$$p(\bar{x}/\bar{k}) = \prod_{t \in T} p_t(x(t)/k(t))$$

Може бути обчислена $p(\bar{k}/\bar{x}) = \dfrac{p(\bar{k}) \cdot p(\bar{x}/\bar{k})}{\sum_{k \in K^T} p(k) \cdot p(\bar{x}/\bar{k})}$ (**)

$\bar{x}$ нехай будем називати зображенням.
=> В цьому обчисленні і полягає саме складність, бо вона залежить від великої кількості розміток
Задача полягає в тому, що для заданого зображення щось сказати про вл-ті прихованої розмітки і знайти розмітку по розподілу.
$\bar{x}^* \mapsto f(\bar{k})$ вл-ть прихованої розмітки.
Це розмітка і генерується за допомогою гіббсового семплеру
Як це зробити?

1) Будь-яким чином нумеруються об'єкти $t_1, t_2, t_3, \dots, t_n$
2) Будь-яким чином вибирається люба розмітка $\bar{k}_0$
3) На основі $\bar{k}_0$ будується нова розмітка $\bar{k}_1$, яка відрізняється від $\bar{k}_0$ тільки в об'єкті t_1, потім будується нова розмітка $\bar{k}_2$, на основі $\bar{k}_1$, яка відрізняється від $\bar{k}_1$ тільки в об'єкті t_2, ..., і будується нова розмітка $\bar{k}_n$ на основі $\bar{k}_{n-1}$, яка відрізняється від $\bar{k}_{n-1}$ в об'єкті t_n

Цей весь процес є одною ітерацією гіббсового семплера, а це крок

$t_1, t_2, \dots, t_i, \dots, t_n$

$\bar{k}_0 \quad \bar{k}_1 \quad \bar{k}_2 \quad \bar{k}_{i-1} \quad \bar{k}_i$

крок

Як реалізується один крок?

На основі $\bar{k}_{i-1}$ будується $\bar{k}_i$. В цій новій розмітці мітки всіх об'єктів, що не є t_i! вони не змінюються.

$k_i(t) = k_{i-1}(t)$, $t \neq t_i$, а $\underset{t=t_i}{k_i(t)}$ вибирається випадково відповідно до якогось розподілу ймовірностей.

Ми знаємо, що мітка в об'єкті t_i залежить від міток в сусідніх об'єктах і від сигналу, що випромінює t_i

$$p\Big(\underset{\substack{\uparrow\\ \text{мітка}\\ \text{в об'єкті } t_i}}{k} / x^*(t_i), k(N(t_i))\Big) = \underset{\substack{\uparrow\\ \text{const}}}{C} \cdot \underset{\substack{\uparrow\\ \text{задано}}}{p_{t_i}}\big(x^*(t_i)/k\big) \cdot \prod_{t' \in N(t_i)} \underbrace{g_{t_i t'}}_{\text{задано.}}\big(k(t_i), k(t')\big)$$

Після достатньо великої кількості ітерацій гіббсового семплера розмітка, що отримується є вип. розміткою, що має цей (**) розподіл ймовірностей.

Рекомендацій щодо кількості ітерацій немає.

11.

$T, K, k: T \to K,\ k \in K^T,\ p: K^T \to \mathbb{R},\ p(k)$

$f_i: K^T \to \mathbb{R},\ i \in I$

$$P \begin{cases} \theta_i = \sum\limits_{k \in K^T} f_i(k) p(k), & i \in I \\ 1 = \sum\limits_{k \in K^T} p(k) \end{cases}$$

$$p^\circ = \underset{p \in P}{\operatorname{argmax}} \left[-\sum_{k \in K^T} p(k) \cdot \log p(k) \right]$$

Розглядаємо випадок, коли θ_i - невідомі, але водночас є вибірка $(k_1, \dots, k_m)$, по якій ми якимось чином хочемо оцінити θ_i°. Використовуємо метод макс. правдоподібності.

$\theta \in \Theta$; $p(x/\theta)$; $(x_1, \dots, x_m) = \bar{X}$

$$p(\bar{X}/\theta) = \prod_{i=1}^{m} p(x_i/\theta)$$

$$\theta^\circ = \underset{\theta \in \Theta}{\operatorname{argmax}}\ p(\bar{x}/\theta) = \underset{\theta \in \Theta}{\operatorname{argmax}} \sum_{i=1}^{m} \log p(x_i/\theta)$$

оцінка для θ

Вводимо вектор-функції

$f: K^T \to \mathbb{R}^n$

$$P_\theta \begin{cases} \theta = \sum\limits_{k \in K^T} f(k) \cdot p(k) \\ 1 = \sum\limits_{k \in K^T} p(k) \end{cases}$$

$$p_\theta = \underset{p \in P_\theta}{\operatorname{argmax}} \left(-\sum_{k \in K^T} p(k) \log p(k) \right)$$

$$P = \{ p_\theta \mid \theta \in \mathbb{R}^n \}$$

p_θ - визначається для конкретного пар-ра θ

P - мн-на p_θ (максим. ентроп. для p_θ $\forall \theta \in$ ...)

Відомо, $p_\theta(k) = C \cdot \exp(\lambda \cdot f(k)) = \dfrac{\exp(\lambda f(k))}{\sum\limits_{k' \in K^T} \exp(\lambda f(k'))}$

Задача:

$x_1, \dots, x_m$

$$\boxed{\theta^* = \underset{\theta \in R^n}{\operatorname{argmax}} \sum_{i=1}^{m} \log p_\theta(x_i)}$$

Можемо передифинулювати

$$p^* = \underset{p \in \mathbf{P}}{\operatorname{argmax}} \sum_{i=1}^{n} \log p(x_i)$$

всі найбільш ентропічні величини вже тут

$$\lambda^* = \underset{\lambda \in R^n}{\operatorname{argmax}} \sum_{i=1}^{m} \log \frac{\exp(\lambda \cdot f(x_i))}{\sum_{\bar{u} \in K^T} \exp(\lambda \cdot f(\bar{u}))}$$

(т.я. $p_\theta(x_i)$ визначена з параметрами $\vec{\lambda}$)

$F(\lambda)$

$$F(\lambda) = \sum_{i=1}^{m} \lambda \cdot f(x_i) - \sum_{i=1}^{m} \log \sum_{\bar{u} \in K^T} \exp(\lambda \cdot f(\bar{u}))$$

$$\frac{dF(\lambda)}{d\lambda^*} = \sum_{i=1}^{m} f(x_i) - \sum_{i=1}^{m} \frac{\sum_{\bar{u} \in K^T} \exp(\lambda^* \cdot f(\bar{u}))\, f(\bar{u})}{\sum_{\bar{u} \in K^T} \exp(\lambda^* \cdot f(\bar{u}))} = 0$$

$p^*(\bar{u})$, θ^*

$$\sum_{i=1}^{m} f(x_i) - \sum_{i=1}^{m} \theta^* = 0$$

$\Downarrow$

$$\boxed{\theta^* = \frac{1}{m} \sum_{i=1}^{m} f(x_i)}$$

12. Найбільш правдоподібне оцінювання параметрів гіббсового поля

T – множина об'єктів, K – множина міток; $\bar{k}: T \to K$ – розмітка $\bar{k} \in K^T$; $p: K^T \to \mathbb{R}$, $p(\bar{k})$

Запишемо експоненційну випадкову величину у векторному вигляді: $f: K^T \to \mathbb{R}^n$; f_i – базисні ф-ії

$$P_\theta \begin{cases} \theta = \sum\limits_{\bar{k} \in K^T} f(\bar{k}) \cdot p(\bar{k}) \\ 1 = \sum\limits_{\bar{k} \in K^T} p(\bar{k}) \end{cases} ; \quad p_\theta = \underset{p \in P_\theta}{\operatorname{argmax}} \left[- \sum_{\bar{k} \in K^T} p(\bar{k}) \cdot \log p(K^T) \right] -$$

– найбільш ентропійна вип. величина.

$P = \{ P_\theta \mid \theta \in \mathbb{R}^n \}$.

Будь-яка із цих величин має такий вигляд:

$$p_\theta(\bar{k}) = C \cdot \exp(\lambda \cdot f(\bar{k})) = \frac{\exp(\lambda \cdot f(\bar{k}))}{\sum\limits_{\bar{k} \in K^T} \exp(\lambda \cdot f(\bar{k}))}$$

Виразимо C, бо $\sum p = 1$ (знаменник)

Запишемо оцінку θ:

$$\boxed{\theta^* = \underset{\theta \in \mathbb{R}^n}{\operatorname{argmax}} \sum_{i=1}^{m} \log p_\theta(\bar{k}_i)}$$

– таким чином p_θ максимізує ймов-ть вибірки $\bar{k}_1, \ldots, \bar{k}_m$

Тобто задача полягає в тому, щоб знайти таке θ і підставити в систему P_θ, щоб ймов-ть p_θ максимізувала ймовірність вибірки $\bar{k}_1, \bar{k}_2, \ldots, \bar{k}_n$

Але замість пошуку θ^* ми можемо сформулювати задачу пошуку такої ф-ї

$$\boxed{p^* = \underset{p \in P}{\operatorname{argmax}} \sum_{i=1}^{m} \log p(\bar{k}_i)}$$

ці функції виглядають таким чином

або можемо ще третє формулювання:

$$\boxed{\lambda^* = \underset{\lambda \in \mathbb{R}^n}{\operatorname{argmax}} \sum_{i=1}^{m} \log \left(\frac{\exp(\lambda f(\bar{k}_i))}{\sum\limits_{\bar{k} \in K^T} \exp(\lambda \cdot f(\bar{k}))} \right)}$$

↑ вектор Лагранжа

Будь-яке із цих формулювань однозначно визначає інше.

Розглянемо задачу λ^*:
Позначимо $\sum\limits_{i=1}^{m} \log \dfrac{\exp(\lambda \cdot f(\bar{k}))}{\sum\limits_{\bar{k} \in K^T} \exp(\lambda \cdot f(\bar{k}))} = F(\lambda)$

$$F(\lambda) = \sum_{i=1}^{m} (\lambda \cdot f(\bar{k})) - \sum_{i=1}^{m} \log \sum_{\bar{k} \in K^T} \exp(\lambda \cdot f(\bar{k}))$$

$$\frac{dF(\lambda)}{d\lambda} = \sum_{i=1}^{m} f(\bar{k}_i) - \sum_{i=1}^{m} \underbrace{\sum_{\bar{k} \in K^T} \underbrace{\frac{\exp(\lambda^* f(\bar{k}))}{\sum_{\bar{u} \in K^T} \exp(\lambda^* f(\ ))}}_{= p^*(\bar{k})} \cdot f(\bar{k})}_{\text{мат очікування } \theta^*} = 0$$

для λ^*

$$\Rightarrow \sum_{i=1}^{m} f(\bar{k}_i) - \sum_{i=1}^{m} \theta^* = 0$$

$$\Rightarrow \theta^* = \frac{1}{m} \sum_{i=1}^{m} f(\bar{k}_i)$$

Перенесемо цей результат на гіббсове вип. поле:

$p(\bar{k})$ – гіббсове вип. поле.

$p_{T'}(s)$ – ймов-ть того, що віконце T' буде заняте фрагментом s.

Відомо $p_{T'}(s) = \sum\limits_{\bar{k} \in K^T} \underbrace{f_{T',s}(\bar{k})} \cdot p(\bar{k})$ – мат очікування.

$\to$ означає проходить розмітка $\bar{k}$ через фрагмент s на віконці T' чи ні
якщо так $\Rightarrow f_{T',s}(\bar{k}) = 1$
$= 0$ інакше.

Задано $\bar{k}_1, \bar{k}_2, \ldots, \bar{k}_m$.

$\Rightarrow$ оцінкою найбільшої правдоподібності для $p_{T'}(s)$ є:

$$p^*_{T'} = \frac{1}{m} \sum_{i=1}^{m} f_{T',s}(\bar{k}_i)$$

Або нехай ми знаємо скільки разів на віконці T' зустрівся фрагмент s
Позначимо це $n_{T'}(s)$

$$\Rightarrow p^*_{T'}(s) = \frac{n_{T'}(s)}{\sum\limits_{s \in K^{T'}} n_{T'}(s)}$$

13.

$X, K, p(x,k), D, W(k,d)$

$$R(q) = \sum_{x\in X} \sum_{k\in K} p_{XK}(x,k)\, W(k, q(x))$$ ← $q(x)$ - детерм. стратегія

дет. стратегія - для кожного вхідного сигналу видає <u>однакове</u> рішення.

рандомізована стратегія - для кожного вхідного сигналу існує розподіл ймовірностей рішень. $q: D \times X \to R$ $(q(d|x))$

$$R(q) = \sum_{x\in X} \sum_{k\in K} p_{XK}(x,k) \sum_{d\in D} \underline{q(d|x)}\, W(k,d)$$

<u>Приклад небайєсової задачі, де q_r краще, ніж $q_{дет}$.</u>

$K \in \{1, 2\}$ p_1, p_2 - апріорні ймов. станів, $p(x|1), p(x|2)$

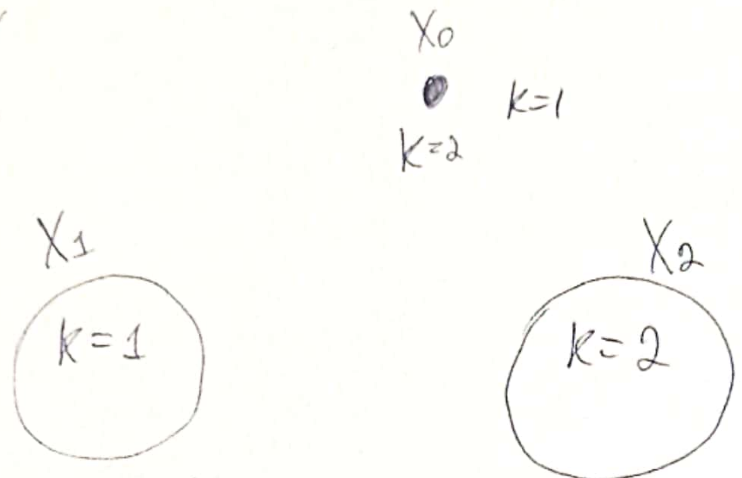
дані обчислюються байєсівським рухом ⇒ помилка розпізнавання. $P_{пом}$

Але існує і інший підхід. На вхід системі подаються не випадково об'єкти, а спочатку подається достатньо велика к-сть об'єктів, про які відомо, що вони знаходяться в стані $k=1$, об'єктів із $k=2$. Оцінка помилки рахується аналогічно, набір об'єктів із $k=2$ окремо. $P_{пом1}$ $P_{пом2}$. Загальна точність $\max\{P_{пом1}, P_{пом2}\}$.

← небайєсівська задача.

Приклад

X



$p(x_0/k=1)=20\%$

$p(x_0/k=2)=30\%$

пом₁ – ймов. помилки розпізнавання об'єктів, які знаходяться в 1 стані

– Розглянемо det стратегію.
Вона на X_1 має давати $k=1$, на $X_2 \to k=2$.
для X_0 має обрати щось одне. $q(x_0)=1 \Rightarrow пом_2 = 30\%$

$q(x_0)=2 \Rightarrow пом_1 = 20\%$

Підсумкова помилка $=30\%$ (для q_{det})

– Розглянемо rand стратегію
на X_1 видає $k=1$, на $X_2 \to k=2$.
а для X_0 – видає ймовірності

$$q(k=1/X_0)=\frac{2}{5}, \quad q(k=2/X_0)=\frac{3}{5} \quad \begin{cases} пом_1 = \frac{3}{5}\cdot 0.2 \cdot 100\% = 12\% \\ пом_2 = \frac{2}{5}\cdot 0.3 \cdot 100\% = 12\% \end{cases}$$

(об'єкт в першому стані, система видає 2 стан)

Підсумкова помилка 12% (для q_{rand})

$12\% < 30\%$

14. Задачі розпізнавання при неповністю відомій статистичній моделі об'єкта розпізнавання. Функція ризику стратегії. Придатні та непридатні стратегії

Досі ми формулювали задачі розпізнавання так:

$x \in X$ – множина сигналів, $k \in K$ – множина станів

(x,k) – випадкова пара; $p(x,k)$ – ймов-ть сигналу у стані k.

$q: X \to K$ – стратегія, яка для кожного x дає стан: $q(x) \in K$

$W: K \times K \to \mathbb{R}$ – ф-я втрат (штрафу).

$W(k,k')$ – визначає штраф, який виплачується, якщо об'єкт знаходиться в стані k', але насправді він знаходиться у стані k.

Якщо об'єкт знаходиться в стані k, а система прийняла рішення $q(x)$ => існує ризик прийняття рішення.

$$R(q) = \sum_{x \in X} p(x) \sum_{k \in K} p(k/x)\, W(k, q(x)) = \sum_{x,k} p(x,k)\, W(k, q(x)) -$$

– детермінований ризик (використовується детермінована стратегія: тобто спостерігаючи сигнал x стратегія вибирає конкретне рішення).

Розглянемо випадок рандомізованої стратегії

$q: K \times X \to \mathbb{R}$;

$q(k/x)$ – якщо система розпізнавання спостерігає x то вона приймає випадкове рішення щодо того, що об'єкт знаходиться у стані k відповідно до розподілу ймов-ей $p(x,k)$.

$$R(q) = \sum_{\substack{x \in X \\ k \in X}} p(x,k) \sum_{k' \in K} q(k'/x)\, W(k,k') \text{ – рандомізований ризик.}$$

цей ризик є лінійною ф-ю стратегії

Вона тут в якості співмножника

На далі використовувать будемо лише рандомізовані стратегії.

Задача з неповністю відомою стат. моделлю

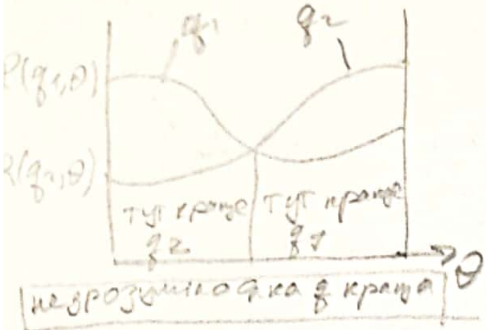
Нехай зараз нам відомо $p(x,k)$ з точністю до якогось параметра $\theta \in \Theta$.

$p(x,k,\theta)$; $\theta \in \Theta$ назвемо моделлю об'єкта.

Θ – множина моделей (кінцева)

=> $\forall$ ранд. ризик залежить не тільки від q, а й від θ

$$\boxed{R(q,\theta) = \sum_{\substack{x \in X \\ k \in K}} p(x,k,\theta) \sum_{k' \in K} q(k'/x) \cdot W(k,k')} \quad (*)$$



Ми маємо зрозуміти як розпізнавати об'єкти, якщо статистична модель об'єкту задана не повністю, а з деякого параметру θ

Цей ризик (*) вказує тепер не просто число, а цілу ф-ію: як ризик залежить від моделі

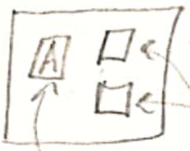
І (*) сам по собі не може сказати, що якась стратегія краща за іншу при певній θ.

Приклад

x - зображення букви, k - ім'я букви

θ - положення букви в полі зору.

Швидше за все: розподіл ймов-стей $p(x,k,\theta)$ на множині імен букви не залежить від θ: тобто ймов-ть появи букви „А" в цьому місці така ж сама як і ймовірність появи „А" в інших місцях



Тобто $p(x,k,\theta) = p(k)\cdot p(x/\theta)$.

Рішення: Якщо задано зобр x^*, то відносно цього зображення приймається рішення, що на цьому зображенні намальована буква з іменем k^*, що вираховується таким чином.

$x^* \mapsto k^* = \arg\min_k \left[\max_\theta p(x^*/k,\theta) \right]$ — ціль: знайти ім'я букви.

⇒ Яке найбільш ймовірне положення букви „А", потім „Б"...?

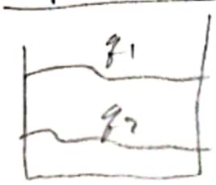
⇒ отримуються числа і вибирається найбільше.

Також можна по іншому задати задачу:

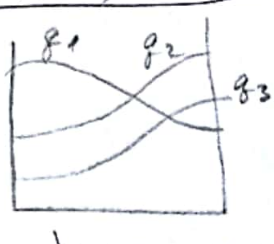
На основі x^* знаходиться не тільки ім'я букви, а й положення.

$x^* \to (k^*,\theta^*) = \arg\max_{(k,\theta)} p(x^*, k/\theta)$ — ціль: знайти ім'я букви і її положення

Стратегія q° наз. непридатною, якщо $\exists q^* \forall\theta \left[R(q^*,\theta) < R(q^\circ,\theta)\right]$



↳ тут q_2 оптимальна



↳ тут q_2 непридатне

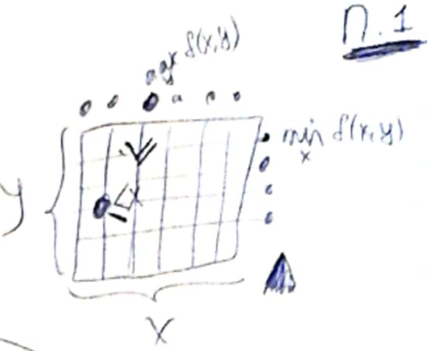
Стратегія q° наз. придатною, якщо $\forall q^* \forall\theta \left[R(q^\circ,\theta) < R(q,\theta)\right]$

15.

Буде потрібна Лема Неймана

$x \in X,\ y \in Y,\ f: X \times Y \to \mathbb{R}$ $\quad f(x,y)$

$$\min_{x\in X}\max_{y\in Y} f(x,y) \geq \max_{y\in Y}\min_{x\in X} f(x,y)$$



В частковому випадку, коли $f(\check{X},\hat{Y})$, то „=“.

(x^0,y^0) – сідлова точка, $\min\limits_{x}\max\limits_{y} f(x,y) = f(x^0,y^0) = \max\limits_{y}\min\limits_{x} f(x,y)$

$x\in X,\ k\in K,\quad p(x,k,\theta),\quad \theta\in\Theta,\quad \theta$ – модель (не обов'язково вип. вел-на)

$q: K\times X \to \mathbb{R},\ q(k|x),\ q\in Q,\quad W(k,k')\in\mathbb{R}$

$$R(q,\theta) = \sum_{x,k} p(x,k;\theta)\sum_{k'\in K} q(k'|x)\,W(k,k')$$

ризик стратегії при моделі θ

Теорема

$q^0\in Q$.

Тоді або $\exists q'\in Q\ \forall\theta\in\Theta\ [R(q',\theta) < R(q^0,\theta)]$ (q^0 – непридатна стратегія)

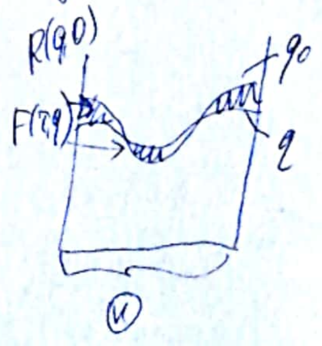
або $\exists \tau:\Theta\to\mathbb{R}: \tau(\theta)\geq 0,\ \sum\limits_{\theta\in\Theta}\tau(\theta)=1: [q^0 = \arg\min\limits_{q\in Q}\sum\limits_{\theta\in\Theta}\tau(\theta)R(q,\theta)]$

$\tau\in T$ $\qquad$ q^0 – байєсівська

Будь яка стратегія є або непридатною або байєсівською, і ні одна стратегія не є байєсівською і непридатною одночасно.

▲ $q^0\in Q$

$$F(\tau,q) = \sum_{\theta\in\Theta}\tau(\theta)\left[R(q,\theta) - R(q^0,\theta)\right]$$



$F(\tau,q)$ – середня величина відмінностями між ризиками стратегій q і q^0.

$F(\tau, q)$ - лінійна ф-я і від q і від τ. $\Rightarrow F(\tilde{\tau}, \tilde{q})$.

$\Rightarrow$ Справедлива лема Неймана.

$$\exists (\tau^\circ, q^\circ): \; q^\circ = \arg\min_{q} F(\tau^\circ, q)$$

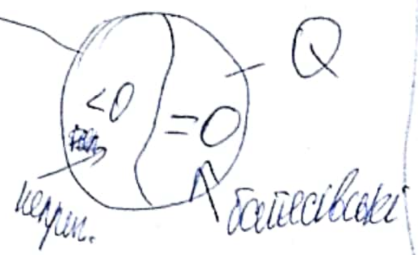
$$\tau^\circ = \arg\max_{\tau} F(\tau, q^\circ)$$

Тв.

$F(\tau^\circ, q^\circ) \le 0$

▲ $\forall \tau \in T \, [F(\tau, q^\circ) = 0] \Rightarrow \forall \tau \in T \, [\min_{q \in Q} F(\tau, q) \le F(\tau, q^\circ) = 0] \Rightarrow$

$\Rightarrow \max_{\tau \in T} \min_{q \in Q} F(\tau, q) \le 0 \Rightarrow F(\tau^\circ, q^\circ) \le 0$ ▲



Тоді можемо поділити Q в залежності від q°.

Доведення теореми буде в тому, що довести ці імплікації.

$F(\tau^\circ, q^\circ) < 0$ $\Updownarrow$ q°-непридатна

$F(\tau^\circ, q^\circ) = 0$ $\Updownarrow$ q°-байєсівська

1. q°-байєсівська $\Rightarrow F(\tau^\circ, q^\circ) = 0$ $\swarrow$ def байєс. стратегія

▲ $\exists \tau^\circ_{\in T} \, \forall q \in Q \, [\sum_{\theta \in \Theta} \tau^\circ(\theta) R(q^\circ, \theta) \le \sum_{\theta \in \Theta} \tau^\circ(\theta) R(q, \theta)] \Rightarrow$

$\Rightarrow \exists \tau^\circ_{\in T} \, \forall q \in Q \, [0 \le \sum_{\theta \in \Theta} \tau^\circ(\theta) (R(q, \theta) - R(q^\circ, \theta))] \Rightarrow$

$\Rightarrow \forall q_{\in Q} \, [F(\tau^\circ, q) \ge 0] \Rightarrow \forall q \in Q \, [\max_{\tau \in T} F(\tau, q) \ge 0 \Rightarrow$

$\Rightarrow \min_{q \in Q} \max_{\tau \in T} F(\tau, q) \ge 0 \Rightarrow F(\tau^\circ, q^\circ) \ge 0$, але $F(\tau^\circ, q^\circ) \le 0 \Rightarrow F(\tau^\circ, q^\circ) = 0$ ▲

15. п 2

2. $F(\tau^\circ, q^\circ) = 0 \Rightarrow q^\circ$ - байєсівська страт.

▲ $\min\limits_{q\in Q} F(\tau^\circ, q) = 0 \Rightarrow \min\limits_{q\in Q} \sum\limits_{\theta\in\Theta} \tau^\circ(\theta)[R(q,\theta) - R(q^\circ,\theta)] = 0 \Rightarrow$

$\Rightarrow \min\limits_{q\in Q} \sum\limits_{\theta\in\Theta} \tau^\circ(\theta) R(q,\theta) - \sum\limits_{\theta\in\Theta} \tau^\circ(\theta) R(q^\circ,\theta) = 0 \Rightarrow$

$\Rightarrow \sum\limits_{\theta\in\Theta} \tau^\circ(\theta) R(q^\circ,\theta) = \min\limits_{q\in Q} \sum\limits_{\theta\in\Theta} \tau^\circ(\theta) R(q,\theta) \Rightarrow q^\circ$ - байєсівська ▲

3. q° - неприн. $\Rightarrow F(\tau^\circ, q^\circ) < 0$

▲ $\exists q^1 \in Q : \forall \theta \in \Theta\, [R(q^1,\theta) - R(q^\circ,\theta) < 0] \Rightarrow$

$\Rightarrow \exists q^1 \in Q \ \max\limits_{\theta\in\Theta} [R(q^1,\theta) - R(q^\circ,\theta)] < 0 \Rightarrow$

$\Rightarrow \exists q^1 \in Q \ \max\limits_{\tau\in T} \sum\limits_{\theta\in\Theta} \tau(\theta)[R(q^1,\theta) - R(q^\circ,\theta)] < 0 \Rightarrow$

$x_i,\ i\in J,$
$\max\limits_{i\in J} x_i$ -?
$\alpha_i \geq 0,\ \sum\limits_{i\in J} \alpha_i = 1$
$\max\limits_{\alpha_i, i\in J} \sum\limits_{i\in J} \alpha_i x_i$ -? Тоді
$\max\limits_{\alpha_i} \sum \alpha_i x_i = \max\limits_{i\in J} x_i$!

$\Rightarrow \min\limits_{q\in Q} \max\limits_{\tau\in T} F(\tau, q) < 0 \Rightarrow F(\tau^\circ, q^\circ) < 0$ ▲

4. $F(\tau^\circ, q^\circ) < 0 \Rightarrow q^\circ$ - неприпадна

$F(\tau^\circ, q^\circ) = \max\limits_{\tau\in T} F(\tau, q^\circ) = \max\limits_{\tau\in T} \sum\limits_{\theta\in\Theta} \tau(\theta)[R(q^1,\theta) - R(q^\circ,\theta)] =$

$= \max\limits_{\theta\in\Theta} [R(q^1,\theta) - R(q^\circ,\theta)] < 0$

$\Downarrow$

$\forall \theta\ [R(q^1,\theta) < R(q^\circ,\theta)]$

$\Rightarrow q^\circ$ - неприпадна ▲ ▲

Scanned with CamScanner

Що з цього слідує?

$p(x,k;\theta)$;

$$k^\circ = \underset{k \in K}{\mathrm{argmax}} \sum_{\theta \in \Theta} \tau(\theta)\, p(x,k;\theta).$$

Якщо ми хочемо потрапити на неприродні стратегії; k° маємо шукати так.

Складність лише в знаходженні $\tau(\theta)$.

16. Привести приклади небайєсовських задач розпізнавання: мінімаксні задачі (minimax deviation)-задачі та інші. Показати, що стратегії рішень цих задач є байєсовими, тобто не є непридатними

$x \in X$ - множина сигналів, що випромінює об'єкт.
$k \in K$ - множина станів об'єкта.
$p(x,k,\theta)$ - ймовірність, визначена до деякого параметра θ (моделі)
$\theta \in \Theta$ - множина моделей.

$q: K \times X \to \mathbb{R}$ - стратегія; $q: x \mapsto \underset{k}{\text{rand}}\ q(k/x)$, $q \in Q$

Запишемо ризик $R(q,\theta)$

$$R(q,\theta) = \sum_{x \in X} \sum_{k \in K} p(x,k,\theta) \cdot \sum_{k' \in K} q(k'/x) w(k,k')$$

Знайдемо $q^* = \underset{q \in Q}{\operatorname{argmin}} \sum_{\theta \in \Theta} \tau(\theta) \cdot R(q,\theta)$

$\tau(\theta) \geq 0$, вага моделі, $\sum_{\theta \in \Theta} \tau(\theta) = 1$; $\tau: \Theta \to \mathbb{R}$.

$\tau \in T$ - множина всіх вагових ф-ій.

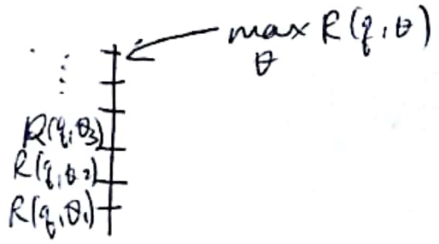
Однозначний вибір стратегії потребує $|K| \cdot |X|$, а однозначний вибір q^* потребує $|\Theta|$ (якщо $|\Theta| < |K| \cdot |X|$)
Потрібно дізнатись як ці ваги не вгадувати, а отримувати автоматично.

<u>Приклад</u>

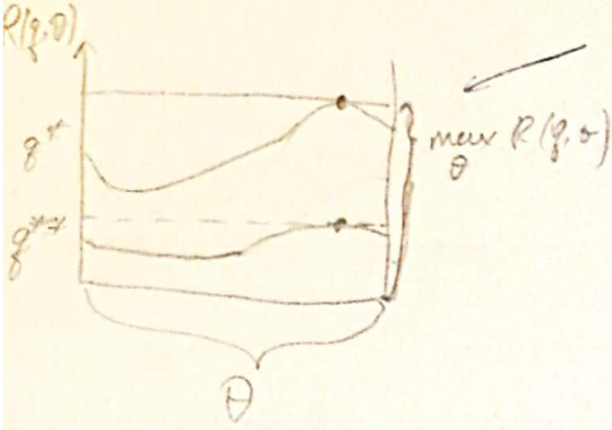
Ми хочемо, щоб наша стратегія така, що
$\forall \theta \in \Theta\ [R(q^*,\theta) \leq C]$: тобто така q^*, яка правильно розпізнає в незалежності від θ

Хочемо знайти таке мінімальне C при якому ця вимога все ще виконується. $R(q,\theta) \leq C$, $\theta \in \Theta$;
Якщо стратегія q задана, то чому має дорівнювати C для виконання нерівності?

$$C = \max_{\theta \in \Theta} R(q,\theta)$$



Потрібно знайти $q^* = \underset{q \in Q}{\operatorname{argmin}} \left[\max_{\theta} R(q,\theta) \right]$ - Важливо те, що ця стратегія отримана таким чином <u>не може бути не придатною</u>



Вона мінімізує максимальний ризик

Нехай q^* не придатна $\Rightarrow \exists\, q^{**}$, що матиме менший ризик за q^*.

$$\Rightarrow \max_\theta R(q^{**},\theta) < \max_\theta R(q^*,\theta) \Rightarrow$$

$\Rightarrow$ це порушує придатність q^*

$\Rightarrow$ ми знаємо, що будь-яка стратегія є байєсівською щодо не придатною

$\Rightarrow q^*$ не є не придатною (є байєсівською).

$$\Rightarrow \exists \tau^* \in T \left[q^* = \arg\min_{q\in Q} \sum_{\theta\in\Theta} \tau^*(\theta) R(q,\theta) \right]$$

$$\tau^* = \arg\max_{\tau\in T} \min_{q\in Q} \sum_{\theta\in\Theta} \tau(\theta) R(q,\theta) = \Phi(\tau)$$

обираємо найгірше τ, що мінімізує стратегію.

В незалежності як виглядають числа $R(q,\theta)$ для $\Phi(\tau)$ завжди увігнута $\cap$ (вона залежить від мін. корд $\tau(\theta)$, що потім мінімізуються).

Алгоритм, що знаходить τ^*

1) $\tau^0 \in T$ Вибираємо навмання τ^0 (початкові ваги).

для $i = 0, 1, \dots$ — нескінченний процес.

На i-му кроці

Для τ^i: $q^i = \arg\min_{q\in Q} \sum_{\theta\in\Theta} \tau^i(\theta) R(q,\theta)$.

застосовуємо стратегію $q^i(x) = \arg\max_{k\in K} \sum_{\theta\in\Theta} \tau^i(\theta) p(x,k,\theta)$.

Для $R(q^i,\theta)$: $\theta^* = \arg\max_{\theta\in\Theta} R(q^i,\theta)$.

Застосовуємо крок субградієнтного методу

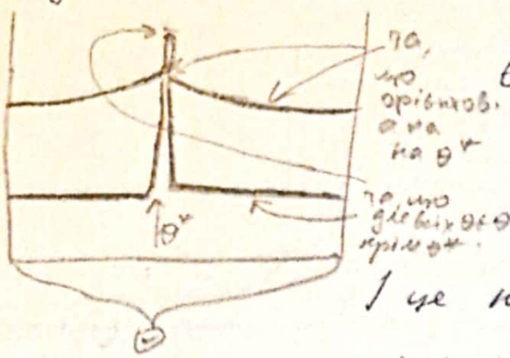
збільшити $\tau(\theta^*)$

зменшити ваги $\tau(\theta)$, $\theta \neq \theta^*$

Вага найгіршої моделі збільшується

Отриманий алгоритм будувався на базі тих знань, що стратегія minmax задачі не може бути не придатною, відповідно вона є байєсовою, відповідно вона мінімізує ризик при деяких вагах. Ті ваги, що беруть там участь є рішенням певної максимізаційної задачі.

Ще один приклад застосування min-max підходу



є деяка стратегія, яка дає досить гарний результат для всіх $\theta \in \Theta$ крім якоїсь особливо неприємної моделі θ^*.

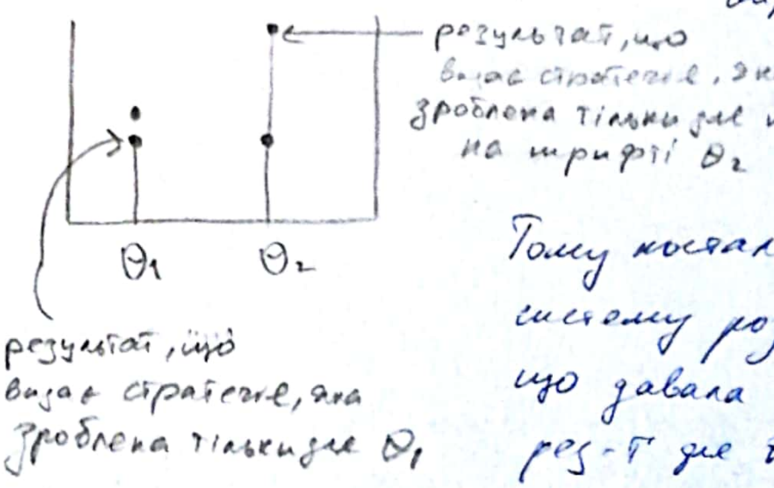
І ця неприємна θ^* така, що навіть стратегія, що намагається мінімізувати ймов-ть помилкового розпізнавання тільки для θ^* все рівно дає поганий результат.

Та стратегія, що орієнтована тільки на θ^* краща, за ту яка видає для всіх $\theta \in \Theta$ крім θ^* гарний рез-т, але вона сильно погано дає рез-т для решти $\theta \in \Theta$

Якщо слідувати принципам min-max, то буде обрана та, що орієнтована на θ^* (бо її максимум нижче за максимум тієї, що для всіх, крім θ^*).

Практичний приклад з шрифтами

Нехай є замовник. Він поставив задачу розробити систему розпізнавання шрифту θ_1. Інженер зробив це, але потім замовник вирішив додати один шрифт θ_2.

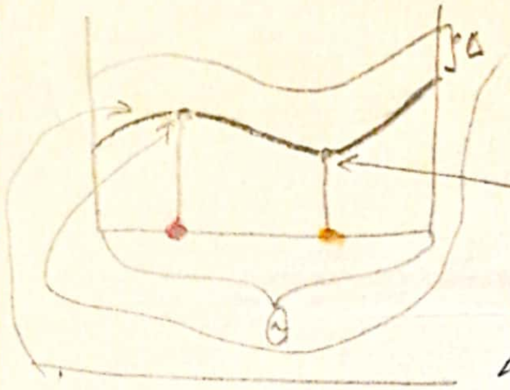


Виявилось, що стратегія, яка дуже добре розпізнавала θ_1 погано розпізнавала θ_2.

Тому постала задача розробити таку систему розпізнавання шрифтів (стратегію) що давала можливо трішки гірший рез-т для θ_1, але задовільний для θ_2.

Приклад minimax deviation strategy.

Нехай маємо таку ситуацію.



Величина ризику стратегії, що мінімізує ймов-ть помилкового розпізнавання для цієї моделі.

Величина ризику стратегії, що мінімізує ймов-ть помилкового розпізнавання, для цієї моделі

Це крива $\min_{q\in Q} R(q,\theta)$

Задача полягає в тому, щоб для заданого проміжку Δ побудувати стратегію ф-я ризику якої повністю лежить у цьому проміжку Δ і при тому, щоб Δ було мінімальним.
Тобто: знайти min Δ при якому виконується.

$$R(q,\theta) - \min_q R(q,\theta) \le \Delta$$

Δ можемо знайти так:

$$\Delta = \max_{\theta\in\Theta}\left[R(q,\theta) - \min_q R(q,\theta)\right]$$ і потрібно знайти

$$q^* = \arg\min_q \max_{\theta\in\Theta}\left[R(q,\theta) - \min_q R(q,\theta)\right]$$

Тими самими судженнями (в попередніх прикладах можна показати, що q^* не є не придатною)

$\Rightarrow$ це q^*: $q^* = \arg\min_q \sum_{\theta\in\Theta} \tau^*(\theta)\left[\quad\right]$

Для цієї ф-ії існує сідлова точка.

певний набір ваг

Цей набір ваг:

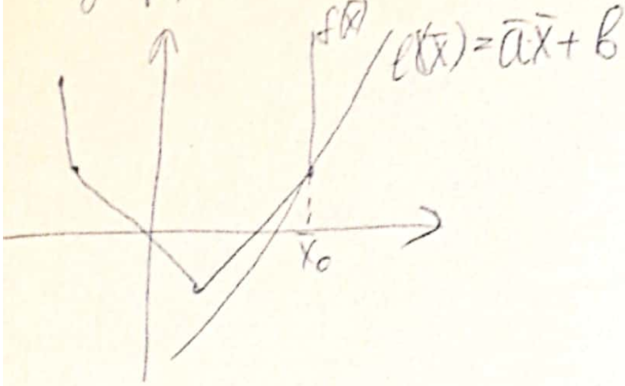
$$\tau^* = \max_{\tau\in T} \min_q \underbrace{\sum_{\theta\in\Theta} \tau(\theta)\left[\quad\right]}$$

⇒ увігнута ф-я ⇒ її максимум може бути знайдений тим чи іншим методом увігнутої максимізації

17.

$f: \mathbb{R}^n \to \mathbb{R}^n$

Поняття супер/субградієнта в точці $\bar{x}_0 \in \mathbb{R}^n$ вводять тоді, коли $f(\bar{x}_0)$ не є диференційовною в $\bar{x}_0$.



1) $f(\bar{x}_0) = l(\bar{x}_0)$

2) $\forall \bar{x} \in \mathbb{R}^n: f(\bar{x}) \geq l(\bar{x})$

$$\nabla_s f(\bar{x}_0) = \nabla l(\bar{x}_0) = \vec{a}.$$

субградієнт ф-ї f в точці $\bar{x}_0$

$$\nabla_s f: \mathbb{R}^n \to \mathbb{R}^n$$

$\bar{x}_0 \in \mathbb{R}^n$, $\exists \nabla f(\bar{x}_0) \Rightarrow \nabla_s f(\bar{x}_0) = \nabla f(\bar{x}_0)$

Властивості субградієнта

1) $\nabla_s(\alpha f(\bar{x})) = \alpha \nabla_s f(\bar{x})$, $\bar{x} \in \mathbb{R}^n$, $\alpha \geq 0$

▲ $g(\bar{x}) = \alpha f(\bar{x})$

$\bar{x}_0$, $\nabla_s f(\bar{x}_0) \Rightarrow l(\bar{x}): \begin{cases} l(\bar{x}_0) = f(\bar{x}_0) \\ l(\bar{x}) \leq f(\bar{x}) \end{cases}$, $x \in \mathbb{R}^n$

$\Downarrow$

$\begin{cases} \alpha l(\bar{x}_0) = \alpha f(\bar{x}_0) = g(\bar{x}_0) \\ \alpha l(\bar{x}) \leq \alpha f(\bar{x}) = g(\bar{x}) \end{cases}$, $x \in \mathbb{R}^n$

$\Downarrow$

$l_1(\bar{x}) = \alpha l(\bar{x})$,

$\nabla_s g(\bar{x}_0) = \nabla l_1(\bar{x}) = \alpha \nabla l(\bar{x}) = \alpha \nabla_s f(\bar{x}_0)$ ▲

---

2) $f_1(\bar{x})$, $f_2(\bar{x})$, $g(\bar{x}) = f_1(\bar{x}) + f_2(\bar{x})$

$\nabla_s g(\bar{x}) = \nabla_s f_1(\bar{x}) + \nabla_s f_2(\bar{x})$

▲ $\bar{x}_0, \nabla_S f_1(\bar{x}_0), \nabla_S f_2(\bar{x}_0)$

$$\begin{cases} l_1(\bar{x}_0) = f_1(\bar{x}_0) \\ l_2(\bar{x}_0) = f_2(\bar{x}_0) \\ l_1(\bar{x}) \le f_1(\bar{x}) \quad \bar{x} \in \mathbb{R}^n \\ l_2(\bar{x}) \le f_2(\bar{x}) \end{cases}$$

$l(\bar{x}) = l_1(\bar{x}) + l_2(\bar{x})$:

$l(\bar{x}_0) = f_1(\bar{x}_0) + f_2(\bar{x}_0) = g(\bar{x}_0)$

$l(\bar{x}) \le f_1(\bar{x}) + f_2(\bar{x}) = g(\bar{x})$, $x \in \mathbb{R}^n$

$\nabla_S g(\bar{x}_0) = \nabla l(\bar{x}_0) = \nabla l_1(\bar{x}_0) + \nabla l_2(\bar{x}_0) = \nabla_S f_1(\bar{x}_0) + \nabla_S f_2(\bar{x}_0)$ ▲

---

3) $f_1(\bar{x}), f_2(\bar{x})$, $g(\bar{x}) = \max\{f_1(\bar{x}), f_2(\bar{x})\}$, $\nabla_S g(\bar{x}_0)$ - ?

▲ $\nabla_S f_1(\bar{x}_0), \nabla_S f_2(\bar{x}_0)$,

$$\begin{cases} l_1(\bar{x}_0) = f_1(\bar{x}_0) \\ l_1(\bar{x}) \le f_1(\bar{x}) \\ l_2(\bar{x}_0) \le f_2(\bar{x}_0) \quad , x \in \mathbb{R}^n \\ l_2(\bar{x}) \le f_2(\bar{x}) \end{cases}$$

- Нехай $f_1(\bar{x}_0) > f_2(\bar{x}_0) \Rightarrow l_1(\bar{x}_0) = f_1(\bar{x}_0) = g(\bar{x}_0)$

$l_1(\bar{x}) \le f_1(x) \le \max\{f_1(\bar{x}), f_2(\bar{x})\} = g(\bar{x}) \Rightarrow l_1(\bar{x}) \le g(\bar{x})$

$\nabla_S g(\bar{x}_0) = \nabla l_1(\bar{x}_0) = \nabla_S f_1(\bar{x}_0)$

- $f_1(\bar{x}_0) \le f_2(\bar{x}_0) \Rightarrow \nabla_S g(\bar{x}_0) = \nabla_S f_2(\bar{x}_0)$ ▲

Узагальнимо: $f_i(\bar{x})$, $i = \overline{1,m}$, $g(\bar{x}) = \max\limits_{i=\overline{1,m}} f_i(\bar{x})$

$\nabla_S g(\bar{x}_0) = \{ k^* = \arg\max\limits_k f_k(\bar{x}_0) \} = \nabla_S f_{k^*}(\bar{x}_0)$

18. Сформулювати процес субградієнтного пошуку і основну його властивість

Нехай $f: D \to \mathbb{R}$. Вектор g наз-ся субградієнтом ф-ії f у точці $x \in D$, якщо $\forall y \in D$ виконується

$$f(y) \geq f(x) + g^T \cdot (y - x)$$

Субдиференціалом f в точці x наз-ся множина всіх субградієнтів f в точці x і позначається $\partial f(x)$

Нехай $f: D \to \mathbb{R}$ - випукла ф-я. Замінюючи в градієнтному методі градієнт на субградієнт отримаємо:

$$x_{k+1} = x_k - \alpha_k g_k, \quad g_k \in \partial f(x) \qquad (1)$$

Головна перевага: застосуємо для $\forall$ випуклої ф-ії

Головний недолік: експоненційна збіжність отримаємо лише в дуже малих випадках. Убільшості збіжність дуже повільна.

Головна проблема: на відміну від градієнтного спуску не можна гарантувати, що $g_k \to 0_n$ навіть, якщо $x_k \xrightarrow[\text{сходиться}]{} x^*$

На далі:

- f - випукла на D ф-я
- f - неперервна по Ліпшицю з константою L, інакше всі субградієнти f рівномірно обмежені на D константою L.
- Відстань від поч. наближення до найближчої точки мінімума обмежено R. $\|x_0 - x^*\| \leq R$

Способи вибору кроку: (ряди).

- константний (постійний): $\alpha_k = \alpha$.
- Розбіжний ряд: $\alpha_k \to 0$, $\sum_{i=0}^{\infty} \alpha_k = \infty$
- Розбіжний ряд зі збіжним рядом квадратів: $\sum_{i=0}^{\infty} \alpha_k = \infty$, $\sum_{i=0}^{\infty} \alpha_k^2 < \infty$
- Нормований: $\alpha_k = \frac{\gamma_k}{\|g_k\|}$; γ_k - одна із вищевказаних послідовностей.

Основна нерівність субградієнтного спуску

$$\varphi_k = \min_{1 \leq i \leq k} f(x_k):$$

$$\|x_{k+1} - x^*\|^2 = \|x_k - x^*\|^2 - 2\alpha_k g_k^T (x_k - x^*) + \alpha_k^2 \|g_k\|^2 \leq$$

$$\leq \|x_k - x^*\|^2 - 2\alpha_k (f(x_k) - f(x^*)) + \alpha_k^2 L^2 \leq$$

$$\leq \|x_0 - x^*\|^2 - 2\sum_{i=0}^{k} \alpha_i (f(x_i) - f(x^*)) + L^2 \sum_{i=0}^{k} \alpha_i^2 =$$

$= \|x_0 - x^*\|^2 - 2(\varphi_k - f(x^*)) \sum_{i=0}^{k} \alpha_i + L^2 \sum_{i=0}^{k} \alpha_i^2$

=> Таким чином:

$$\varphi_k - f(x^*) \le \frac{R + L^2 \sum_{i=1}^{k} \alpha_i^2}{2 \sum_{i=1}^{k} \alpha_i} \qquad (2)$$

Теорема (про збіжність субградієнтного спуску).
Якщо f - випукла на D ф-я, x^* - точка мінімума f на D, f - неперервна по Ліпшицю з константою L, $\|x_0 - x^*\| < R$ => для найкращого наближення по (1):

- При $\alpha_k = \alpha$: $\overline{\lim_{k \to +\infty}} \varphi_k - f(x^*) \le \frac{\alpha L^2}{2}$.
- При $\alpha_k \to 0$. $\sum_{k=1}^{\infty} \alpha_k = \infty$: $\varphi_k - f(x^*) = 0$

▲ Перше твердження виводиться безпосередньо підстановкою $\alpha_k = \alpha$ і граничним переходом:
Для другого в силу $\alpha_i \to 0$ існує $N_1 : \forall k > N_1 \ a_i < \frac{\varepsilon}{L^2}$, а в силу $\sum_{i=0}^{\infty} a_i = \infty$ існує $N_2 : \forall n > N_2 \ \sum_{i=0}^{n} \alpha_i > \frac{R}{\varepsilon}$ таким чином для $k \geq \max\{N_1, N_2\}$ отримаємо:

$$\varphi_k - f(x^*) \le \frac{R + L^2 \sum_{i=1}^{k} \alpha_i^2}{2 \sum_{i=0}^{k} \alpha_i} \le \frac{R}{2 \sum_{i=0}^{k} \alpha_i} + \frac{L^2 \cdot \varepsilon \cdot \sum_{i=0}^{k} \alpha_i}{2 \cdot \sum_{i=0}^{k} \alpha_i} <$$

$$< \frac{R}{2 \cdot \frac{R}{\varepsilon}} \cdot \frac{\varepsilon \cdot \sum_{i=0}^{k} \alpha_i}{2 \cdot \sum_{i=0}^{k} \alpha_i} = \varepsilon$$

Оцінка (2) є ф-я змінних $\alpha_1, \ldots, \alpha_k$, мінімізація якої буде давати гарантію ~~найкра~~ кращої збіжності.
Позначимо: $\Phi(\alpha_1, \ldots, \alpha_k) = \frac{R + L^2 \sum_{i=1}^{k} \alpha_i^2}{2 \sum_{i=1}^{k} \alpha_i}$

Диференціюючи по α_i:

$$\frac{\partial \Phi}{\partial \alpha_i} = \frac{4 L^2 \alpha_i \sum \alpha_i - 2(R + L^2 \sum \alpha_i^2)}{(2 \sum \alpha_i)^2} = 0$$

Рівняння ідентичне для різних i, що дає рівність всіх α_i.
Використовуючи це отримаємо: $4 L^2 k \alpha_i^2 = 2R + 2L^2 k \alpha_i^2$

$$\Rightarrow \alpha_i = \frac{\sqrt{R}}{L \sqrt{k}}$$

19.

$x \in X,\ k \in K,\ p(x,k|\theta),\ \theta \in \Theta$; $q(k|x)$ - ? , $R(q,\theta)$

q^θ $R^0(\theta)$ – вимоги до точності системи ($\min_{q\in Q} R(q,\theta)$)

$q^0 = \arg\min_{q\in Q} \max_{\theta\in\Theta} [R(q,\theta) - R^0(\theta)]$, але знаємо, що

$\exists\ \tau^0 \in T:\ q^0 = \arg\min_{q\in Q} \sum_{\theta\in\Theta} \tau^0(\theta) [R(q,\theta) - R^0(\theta)]$

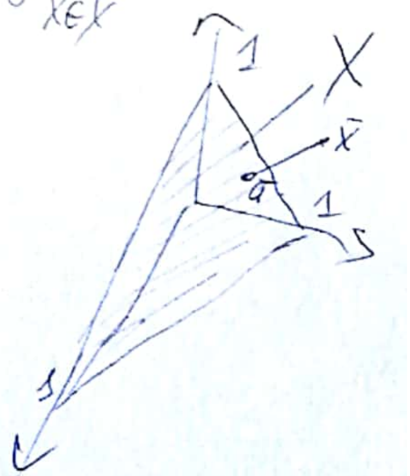
$\tau^0 = \arg\max_{\tau\in T} \min_{q\in Q} \sum_{\theta\in\Theta} \tau(\theta) [R(q,\theta) - R^0(\theta)]$

Випукла задача оптимізації з обмеженнями ($\sum_{\theta\in\Theta} \tau(\theta) = 1$, $\tau(\theta) \geq 0\ \forall \theta\in\Theta$)

Як робити оптимізацію з обмеженнями?

$\bar{x} = (x_1, \dots, x_i, \dots, x_n) \in \mathbb{R}^n$, $f: \mathbb{R}^n \to \mathbb{R}$, $g: \mathbb{R}^n \to \mathbb{R}^n$, g - субград. $f(x)$

$\max_{x\in X} f(x)$ - ? $X = \{\bar{x} \mid x_i \geq 0,\ \sum_{i=1}^{n} x_i = 1\}$; $\bar{x}^0 = \arg\max_{\bar{x}\in X} f(x)$



$\Pi p_X(\bar{x}) = \arg\min_{a\in X} \|\bar{x} - a\|^2$

$\gamma_i \xrightarrow[i\to\infty]{} 0$

$\sum_{i=1}^{\infty} \gamma_i \Rightarrow \infty$

γ_i - крок

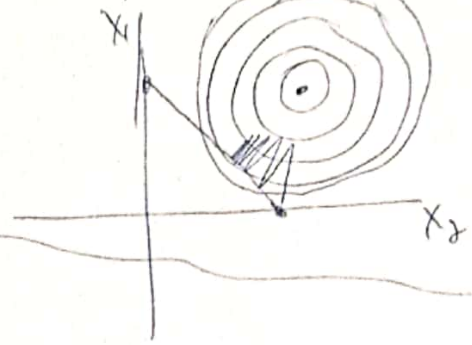
$\bar{x}_0, \bar{x}_1, \dots, \bar{x}_i, x_{i+1}, \dots$

$\bar{x}_0 \in X$ (для спрощення обчислень)

Нехай уже побудовано $\bar{x}^i$.

$$\bar{x}^i \mapsto \bar{x}'_{i+1} = \bar{x}_i + \gamma_i \cdot g(\bar{x}_i) \longmapsto \bar{x}_{i+1} = \Pi p_X(\bar{x}'_{i+1}).$$

Повторюємо необхідну к-сть разів.



$$\lim_{i\to\infty} f(\bar{x}^i) = \max_{\bar{x}\in X} f(\bar{x})$$

Застосуємо для отримання τ^*:

$$\begin{cases} q^* = \arg\min\limits_{q\in Q} \max\limits_{\theta\in\Theta} [R(q,\theta) - R^0(\theta)] \\ \tau^* = \arg\max\limits_{\tau\in T} \underbrace{\min\limits_{q\in Q} \sum\limits_{\theta\in\Theta} \tau(\theta)(R(q,\theta) - R^0(\theta))]}_{-F(\tau)} \end{cases}$$

На i-му кроці маємо $\tau^i(\theta)$

1. $q^i = \arg\min\limits_{q\in Q} \sum\limits_{\theta\in\Theta} \tau^i(\theta)(R(q,\theta) - R^0(\theta))$, обчислюємо $R(q^i;\theta)$

$$F(\tau) = \sum_{\theta\in\Theta} \tau(\theta)[R(q^i;\theta) - R^0(\theta)]$$

$|\Theta|$-вимірний вектор $g(\tau^i)$ складається з ел-тів

$$\frac{\partial F(\tau)}{\partial \tau(\theta)} = R(q^i;\theta) - R^0(\theta)$$

$$\tau'(\theta) = \tau^i(\theta) + \gamma_i [R(q^i;\theta) - R^0(\theta)],$$

$$\tau^{i+1} = \Pi p_T(\tau')$$

Таким чином на певному кроці отримуємо ваги $\tau(\theta)$ для безпосереднього обчислення q^*. Стратегія, отримана таким чином ні в якому разі не може бути непридатною.