МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАВЧАЛЬНО-НАУКОВИЙ КОМПЛЕКС "ІНСТИТУТ ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ"
НАЦІОНАЛЬНОГО ТЕХНІЧНОГО УНІВЕРСИТЕТУ УКРАЇНИ “КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ”
КАФЕДРА МАТЕМАТИЧНИХ МЕТОДІВ СИСТЕМНОГО АНАЛІЗУ

Дипломна робота

Деякі застосування байєсівських методів у машинному навчанні

Виконав: Бондаренко Є.С., КА-21
Науковий керівник: к. ф.-м. н., доц. Каніовська І. Ю.

Дипломна робота

- **Об’єкт дослідження**: алгоритми машинного навчання.
- **Предмет дослідження**: байєсівський статистичний підхід та його застосування в задачах машинного навчання.
- **Мета роботи**: проаналізувати предмет дослідження та дослідити можливість покращення результативності деяких алгоритмів машинного навчання за допомогою байєсівських методів.
- **Метод дослідження**: методи штучного інтелекту, теорії ймовірностей, статистичного аналізу, лінійної алгебри, дискретної математики, оптимізації (чисельні та аналітичні).

Актуальність дослідження

- Сьогодні, у добу big data, алгоритми **машинного навчання** є надзвичайно популярними:
  - біоінформатика
  - комп’ютерний зір
  - розпізнавання образів
  - обробка природньої мови
  - пошукові двигуни
  - рекомендаційні системи

- Такі IT-гіганти як **Google**, **Microsoft**, **Facebook** та багато інших вкладають щорічно багато коштів у дослідження в області машинного навчання.

- У свою чергу, **байєсівські методи** дозволяють вирішити велику кількість проблем та покращити результативність алгоритмів машинного навчання.

Постановка завдання до дипломної роботи

- **Задача 1**. Відновлення лінійної регресії методами байєсівського аналізу (в залежності від апріорних знань) та порівняння з класичним алгоритмом МНК.
- **Задача 2**. Класифікація текстів за допомогою Naïve Bayes classifier та Maximum entropy classifier та порівняння із класичним алгоритмом Support vector machine (SVM).
- **Задача 3**. Автоматичне знаходження розмірності підпростору в задачі зменшення розмірності даних за допомогою Байєсівського PCA та порівняння з класичним алгоритмом PCA.

Задача 1. Відновлення лінійної регресії

- **Постановка задачі**: Нехай є дані, отримані в результаті гіпотетичного експерименту, які лінійно залежать від деякої величини X. Дані надходять з шумом, дисперсія якого невідома. Необхідно знайти коефіцієнти лінійної залежності.

- **Вхідні дані**:

$$x_1 \dots x_m \in [0; 10], \qquad m = 20$$

$$y = a \cdot x + b + \varepsilon, \qquad a = 8.7, b = 4.4$$

ε – гаусівський стандартний білий шум

Задача 1. Байєсівський аналіз

- Функція правдоподібності:

$$f(X|a,b,\sigma) = \prod_{i=1}^{m} \mathcal{N}(a \cdot x_i + b, \sigma^2), \sigma > 0$$

- Апріорні розподіли параметрів:

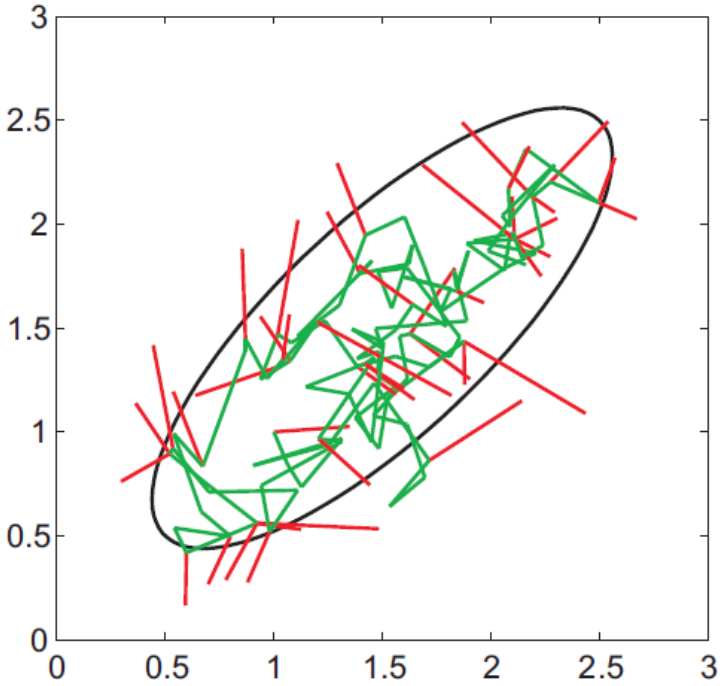
$a \sim U(\Delta_a), b \sim U(\Delta_b), \sigma \sim U(\Delta_\sigma)$ для деяких $\Delta_a, \Delta_b \subset \mathbb{R}$ та $\Delta_\sigma \subset \mathbb{R}_+$

- Оцінки $\hat{a} = \mathbb{E}_p[a]$, $\hat{b} = \mathbb{E}_p[b]$, де:

$$p(a,b,\sigma|X) \propto f(X|a,b,\sigma) \cdot \pi(a,b,\sigma) = f(X|a,b,\sigma) \cdot \pi(a) \cdot \pi(b) \cdot \pi(\sigma)$$

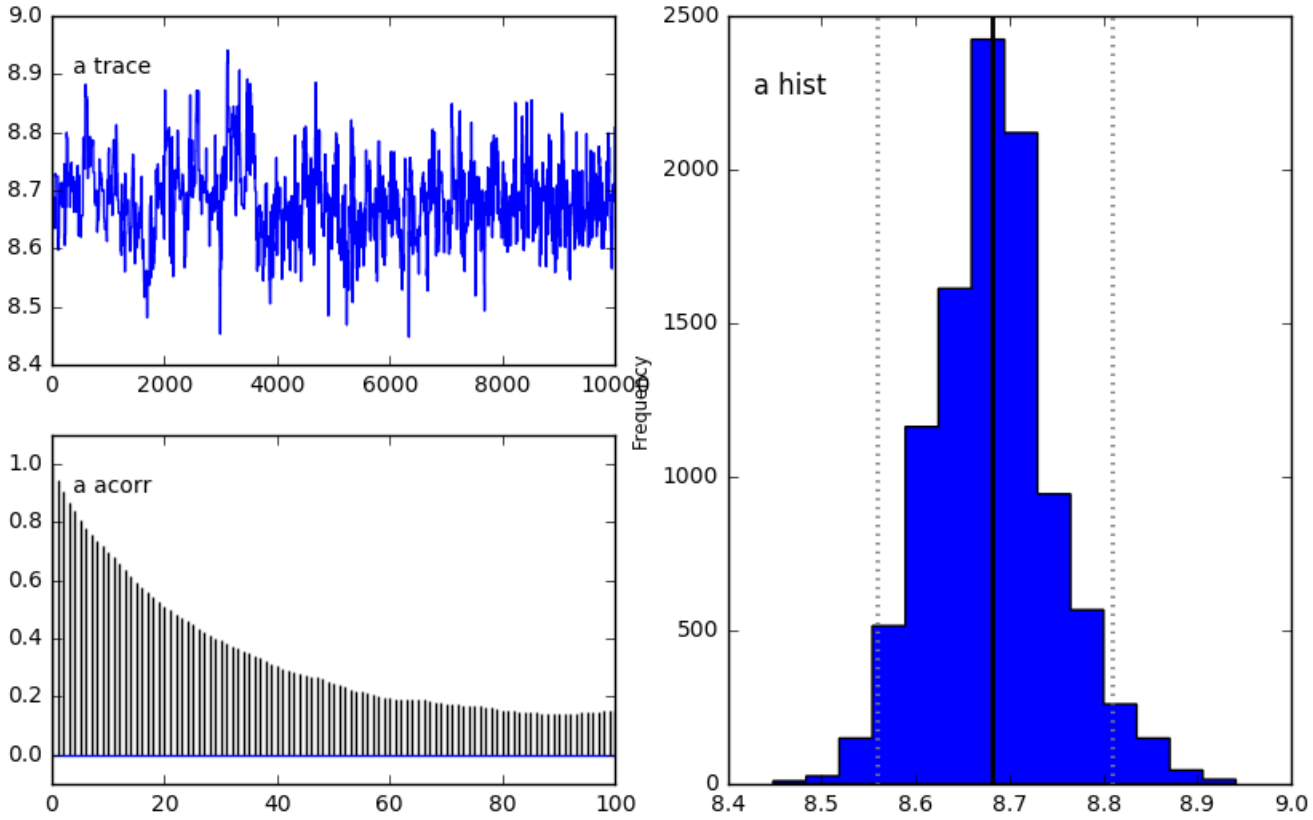
Задача 1. Моделювання апостеріорного розподілу (MCMC)

MCMC за схемою Метрополіса-Хастінгса:



Параметри:

- кількість ітерацій: 15000;
- період розігріву: 5000.



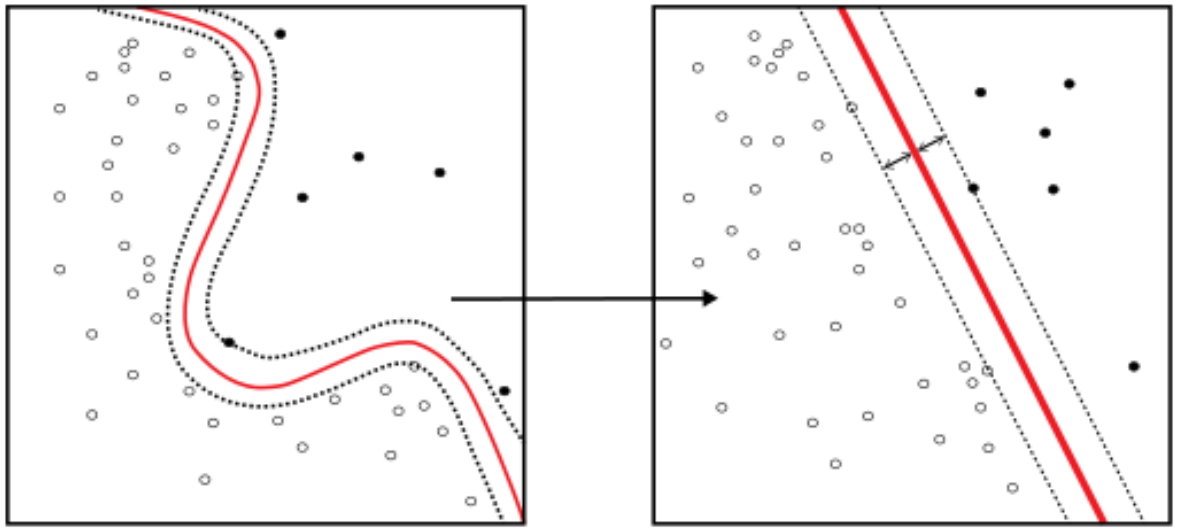
Задача 1. Резултати

$\hat{a}_{LSM}$	$\hat{b}_{LSM}$	RMSE
8.6781	4.7776	0.2761

a, b	σ	$\hat{a}_{PM}$	$\hat{b}_{PM}$	RMSE
$\sim U(-10^4, 10^4)$	$\sim U(0, 10^5)$	8.6081	5.1004	0.3688
	$\sim U(0, 10^3)$	**8.7295**	**4.4979**	**0.2609**
	$\sim U(0, 10)$	8.6619	4.8514	0.2855
	$\sim U(0, 2)$	8.6650	4.9426	0.3827
$\sim U(0, 100)$	$\sim U(0, 10^5)$	8.7224	4.6078	0.3267
	$\sim U(0, 10^3)$	**8.6851**	**4.6268**	**0.1589**
	$\sim U(0, 10)$	**8.6551**	**4.7836**	**0.2096**
	$\sim U(0, 2)$	**8.7026**	**4.5636**	**0.1768**
$\sim U(0, 10)$	$\sim U(0, 10^5)$	**8.7189**	**4.4796**	**0.1830**
	$\sim U(0, 10^3)$	**8.6818**	**4.7015**	**0.2176**
	$\sim U(0, 10)$	**8.6992**	**4.5957**	**0.1917**
	$\sim U(0, 2)$	**8.6765**	**4.6193**	**0.1244**

Задача 2. Класифікація текстів

- Класифікація текстів:
  - спам-фільтри
  - класифікація пошти на групи
  - автоматична ідентифікація мови, жанру, настрою автора та ін.
- До сих пір не існує універсального алгоритму. Один з лідерів – SVM. Головні недоліки:
  - лінійність;
  - складно інтерпретувати деякі параметри;
  - модель не ймовірнісна, складно оцінити $\mathbb{P}(y_i|\boldsymbol{x})$
- Будуть розглянуті Naïve Bayes classifier та Maximum entropy classifier



Задача 2. Класифікація текстів

- **Постановка задачі**: Нехай є множина документів D та множина категорій C. Нехай також є тренувальний промаркований набір пар «документ-категорія» $(d, c), d \in D, c \in C$. Необхідно запропонувати алгоритм, який на основі цих даних міг би віднести нові документи до однієї чи декількох категорій з C.
- **Вхідні дані**: відкритий датасет Reuters – 9517 уривків новин за 90 можливими категоріями англійською мовою.
- Побудуємо три бінарних класифікатори для розпізнавання трьох категорій з найбільшою кількістю документів в датасеті – earn, acq та money-fx.

Задача 2. Naïve Bayes classifier

- Моделюється умовний розподіл $\mathbb{P}(c|\boldsymbol{d})$ – категорії c за умови що спостерігається документ $\boldsymbol{d}$:

$$\mathbb{P}(c|\boldsymbol{d}) = \frac{\mathbb{P}(c) \cdot \mathbb{P}(\boldsymbol{d}|c)}{\mathbb{P}(\boldsymbol{d})} \propto \mathbb{P}(c) \cdot \mathbb{P}(\boldsymbol{d}|c)$$

- Припущення умовної незалежності ознак:

$$\forall i \neq j \;\; \forall c \in C \;\; \mathbb{P}\big(d_i, d_j|c\big) = \mathbb{P}(d_i|c) \cdot \mathbb{P}\big(d_j|c\big) \quad \Longrightarrow \quad \mathbb{P}(\boldsymbol{d}|c) = \prod_{i=1}^{n} \mathbb{P}(d_i|c)$$

- Змоделювавши розподіл, клас обирається за оцінкою MAP:

$$\hat{c} = \underset{c \epsilon C}{\operatorname{argmax}} \left\{ \mathbb{P}(c) \cdot \prod_{i=1}^{n} \mathbb{P}(d_i|c) \right\}$$

Задача 2. Принцип максимальної ентропії

- Нехай відома деяка інформація (деякі обмеження) щодо апріорного розподілу.

- Тоді слід обирати апріорний розподіл, який за даних обмежень максимізує ентропію:

$$\mathbb{P}\{X = x_k\} = p_k,\ k = 1,2, \ldots \qquad f(x)$$

$$H(X) = -\sum_k p_k \cdot \log(p_k) \qquad H(X) = -\int_{-\infty}^{+\infty} f(x) \log\big(f(x)\big)\, dx$$

Задача 2. Maximum entropy classifier

- Моделюється $\mathbb{P}(c|d)$.
- $f_i(d, c) \in \mathbb{R}$ – набір ознак документів. Обмеження:

$$\forall i \quad \frac{1}{|D|}\sum_{d \in D} f_i\big(d, c(d)\big) = \mathbb{E}_{c|d}[f_i(d, c)]$$

- Можна показати, що за цих обмежень розподіл матиме наступну форму:

$$\mathbb{P}_{\boldsymbol{\Lambda}}(c|d) \propto \exp\left(\sum_i \lambda_i f_i(d, c)\right),$$

λ_i – параметри, які необхідно оцінити.

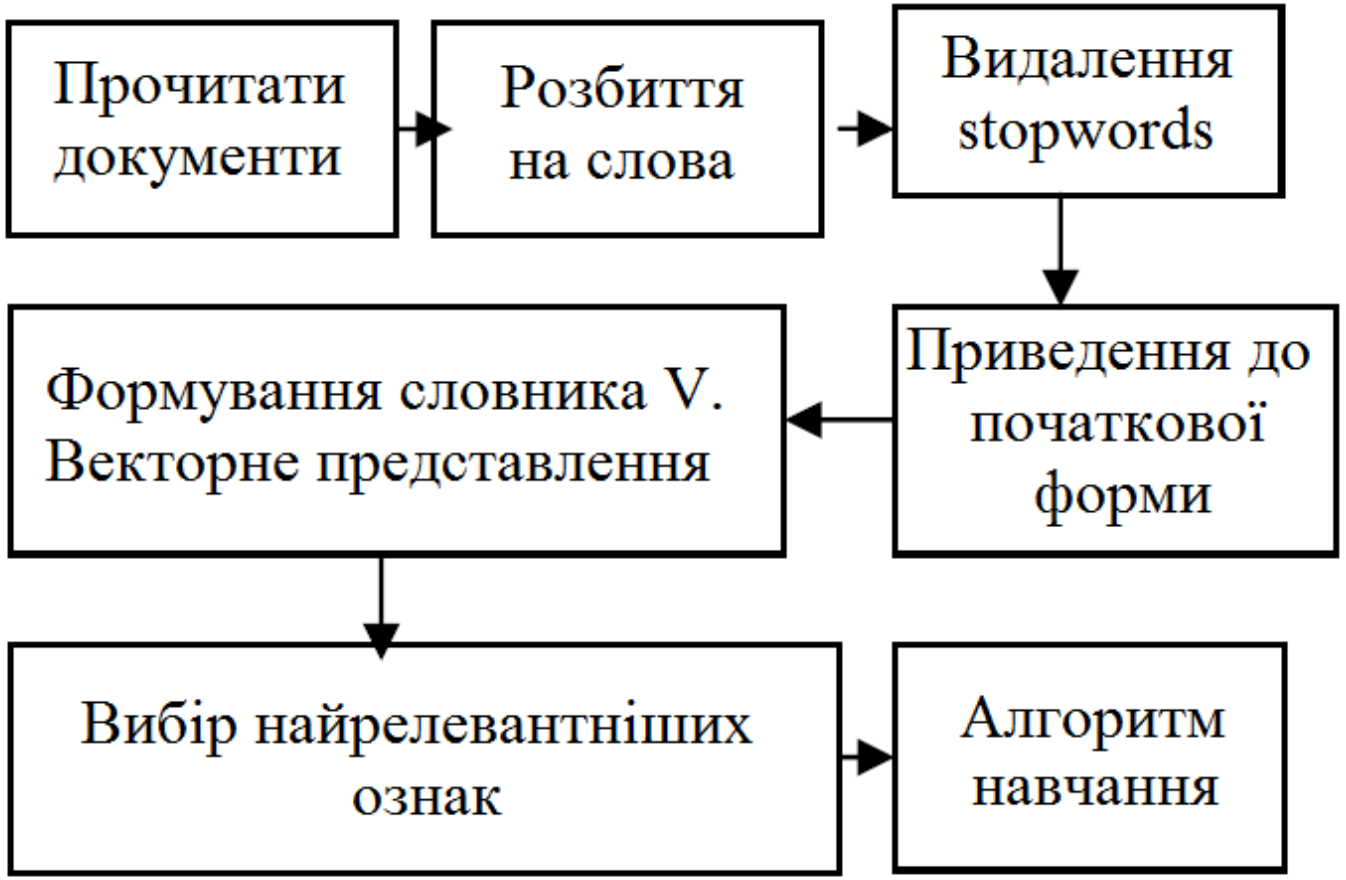
Задача 2. Maximum entropy classifier

- Можна показати, що задача максимізації ентропії в даному випадку еквівалентна максимізації (логарифмічної) правдоподібності:

$$\ell(\mathbf{\Lambda}|D) = \log\left(\prod_{d\in D} \mathbb{P}_{\mathbf{\Lambda}}(c(d)|d)\right) = \sum_{d\in D} \log \sum_{c\in C} \exp \sum_{i} \lambda_i f_i(d, c(d)) \to \max_{\mathbf{\Lambda}}$$

- Для максимізації $\ell(\mathbf{\Lambda}|D)$ використано ітераційний алгоритм Improved Iterative Scaling (IIS), ідея якого базується на опуклості вгору даної функції. Умова зупинки алгоритму $|\ell_{n+1} - \ell_n| < 10^{-6}$.

Задача 2. Алгоритм виділення ознак з документів



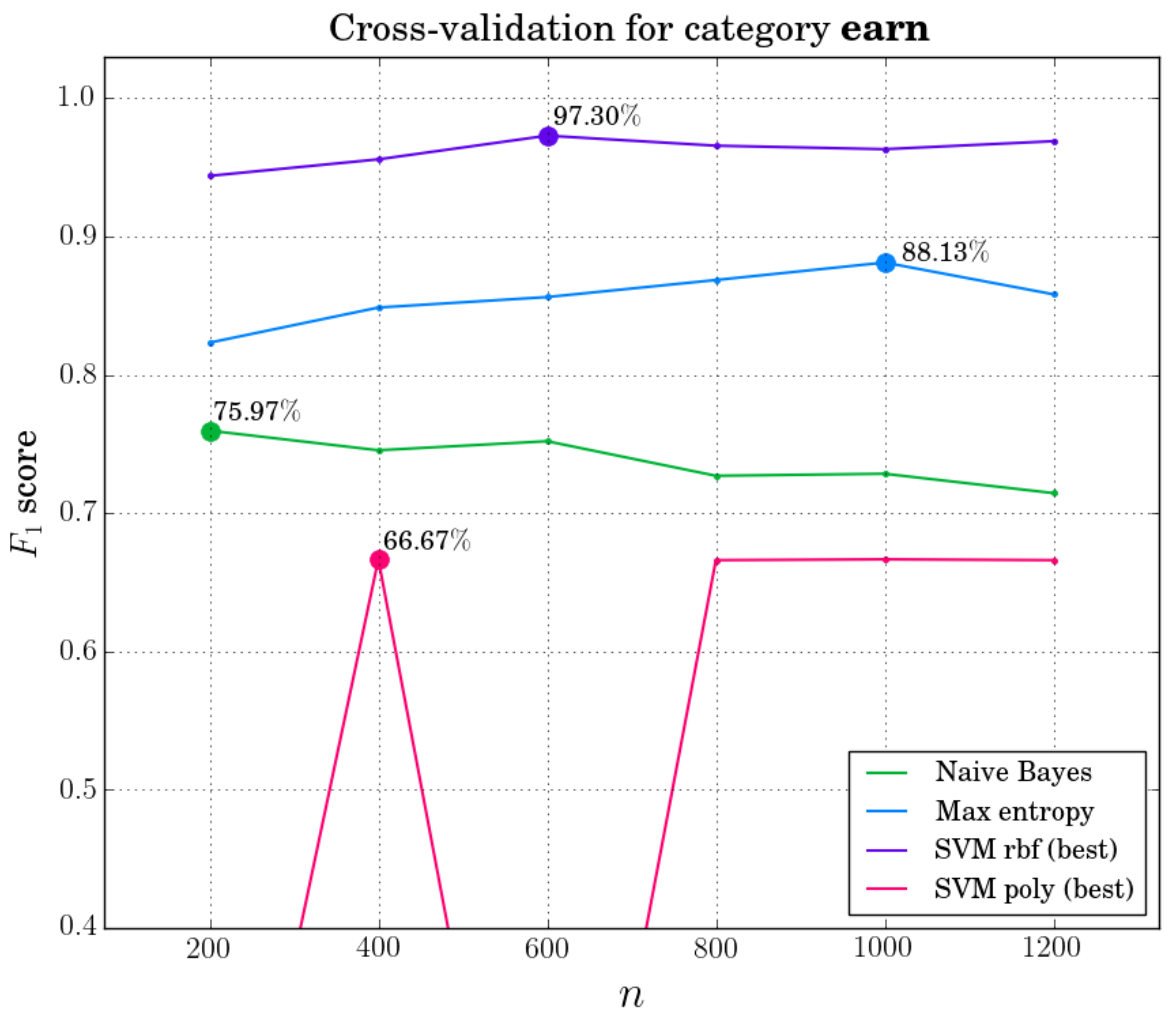
Задача 2. Постановка експерименту

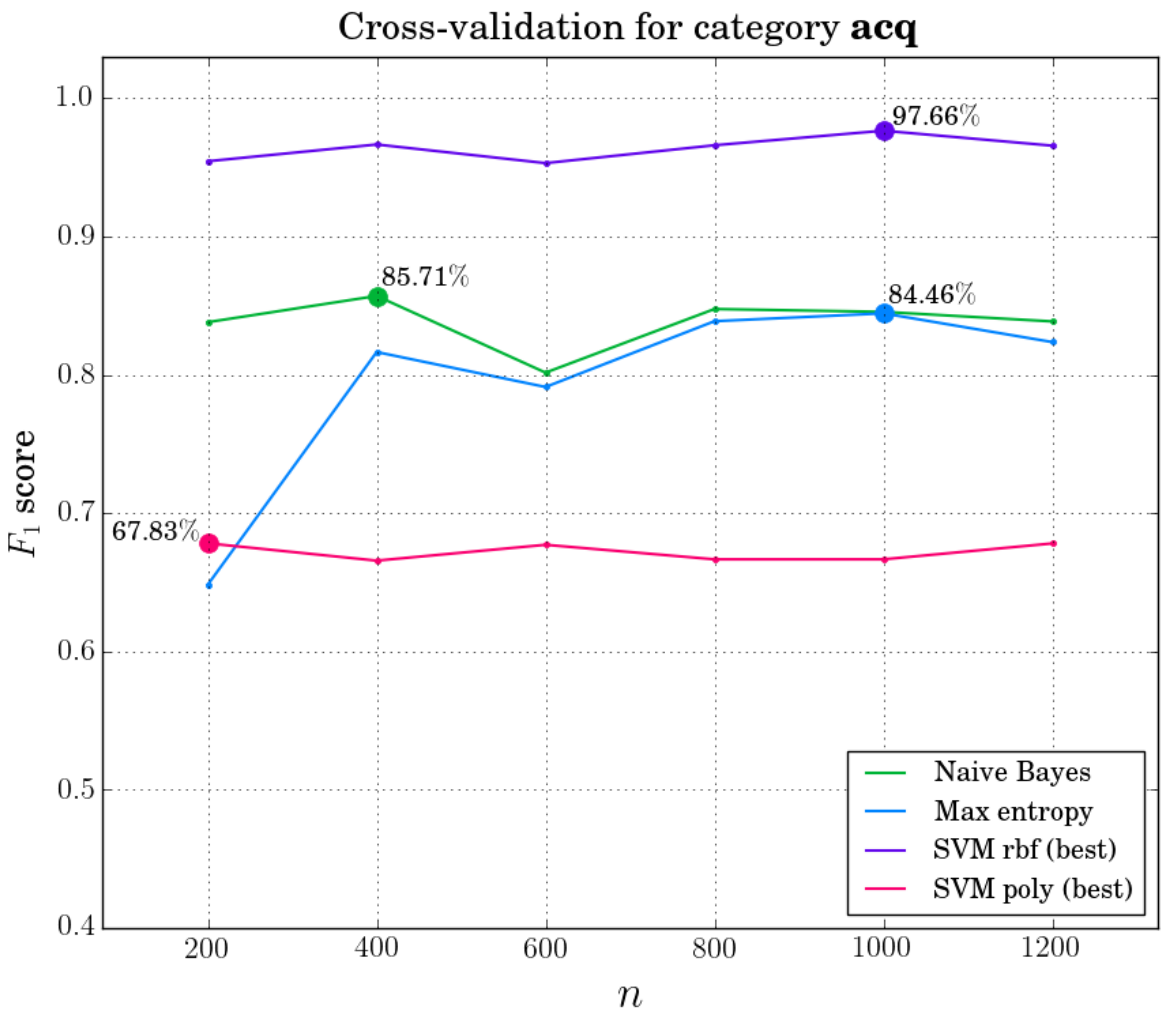
- Кількість ознак $n \in \{200, 400, 600, 800, 1000, 1200\}$.
- SVM запускався з наступними ядрами:
  - гаусівське (rbf): $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\gamma \cdot \left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2\right)$ для $\gamma \in \{0.1, 0.5, 1.0, 2.0, 5.0, 10, 20, 50\}$;
  - поліноміальне: $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(\boldsymbol{x}_i \cdot \boldsymbol{x}_j\right)^d$ для $d \in \{1, 2, 3, 4, 5\}$.
- Вибірка = Тренувальна (80%) : cross-validation (10%) : тестова (10%)
- Метрика – оцінка F_1:

$$F_1 = \frac{2PR}{P+R}, P(precision) = \frac{\text{кількість документів, що правильно передбачені як 1}}{\text{кількість документів, що передбачені як 1}},$$

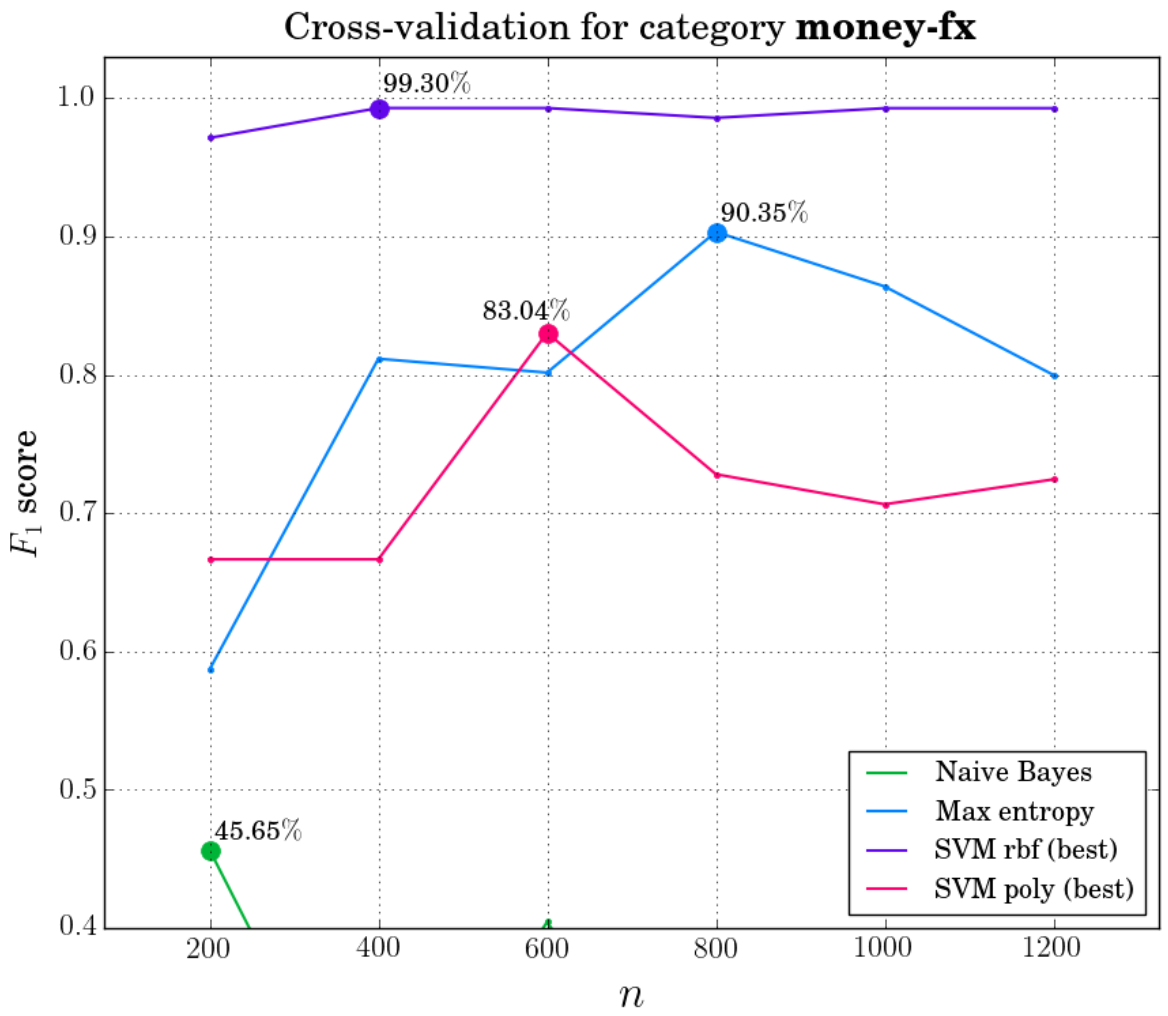
$$R(recall) = \frac{\text{кількість документів, що правильно передбачені як 1}}{\text{кількість документів, що дійсно належать до 1}}$$

Задача 2. Резултати





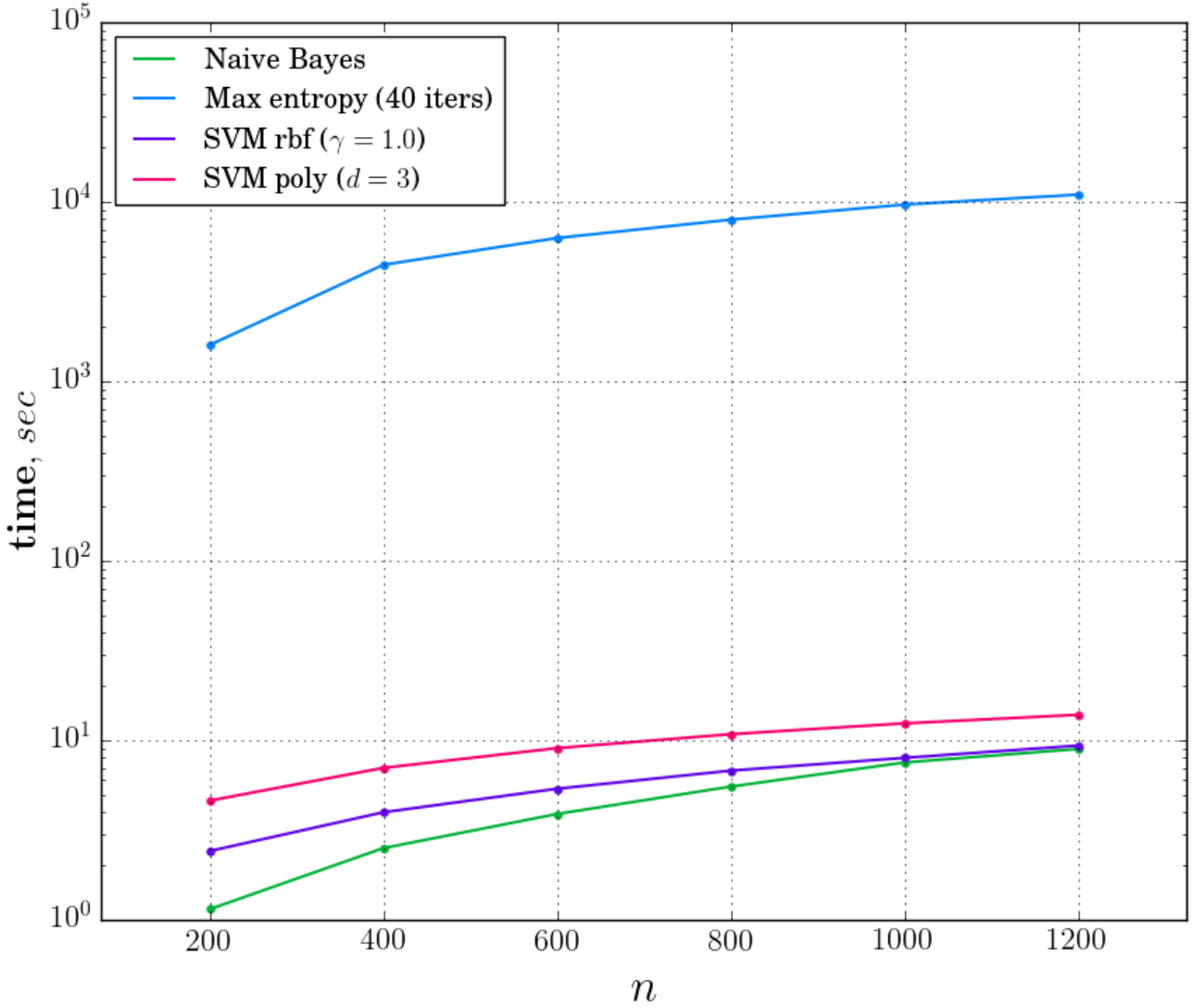
Задача 2. Результати



Оцінки F_1 для тестової вибірки для кращих моделей з cross-validation

Категорія	NBC	MEC	SVM
earn	77.56%	88.86%	95.95%
acq	81.55%	82.92%	95.10%
money-fx	42.22%	86.27%	97.90%

Задача 2. Час роботи алгоритмів



Задача 3. Зменшення розмірності даних

- Задача зменшення розмірності даних:
  - зменшення обчислювальних витрат
  - стиснення даних для ефективного їх зберігання
  - візуалізація даних
  - виявлення нових (прихованих) ознак в даних та ін.
- Класичний алгоритм – PCA, що полягає у пошуку ортонабору векторів $\boldsymbol{w}_1 \ldots \boldsymbol{w}_d$ в просторі $\mathbb{R}^n$ $(d < n)$ – *головних компонент*, що дисперсія спроектованих даних на їх лінійну оболонку буде максимальною $\Leftrightarrow$ середньоквадратична похибка реконструкції буде мінімальною. Головні недоліки:
  - до уваги беруться тільки середнє та коваріація
  - міра інформативності – дисперсія
  - лінійність
  - трудомісткі операції
  - необхідність задання параметра d
- Будуть розглянуті ймовірнісні моделі PCA.

Задача 3. Зменшення розмірності даних

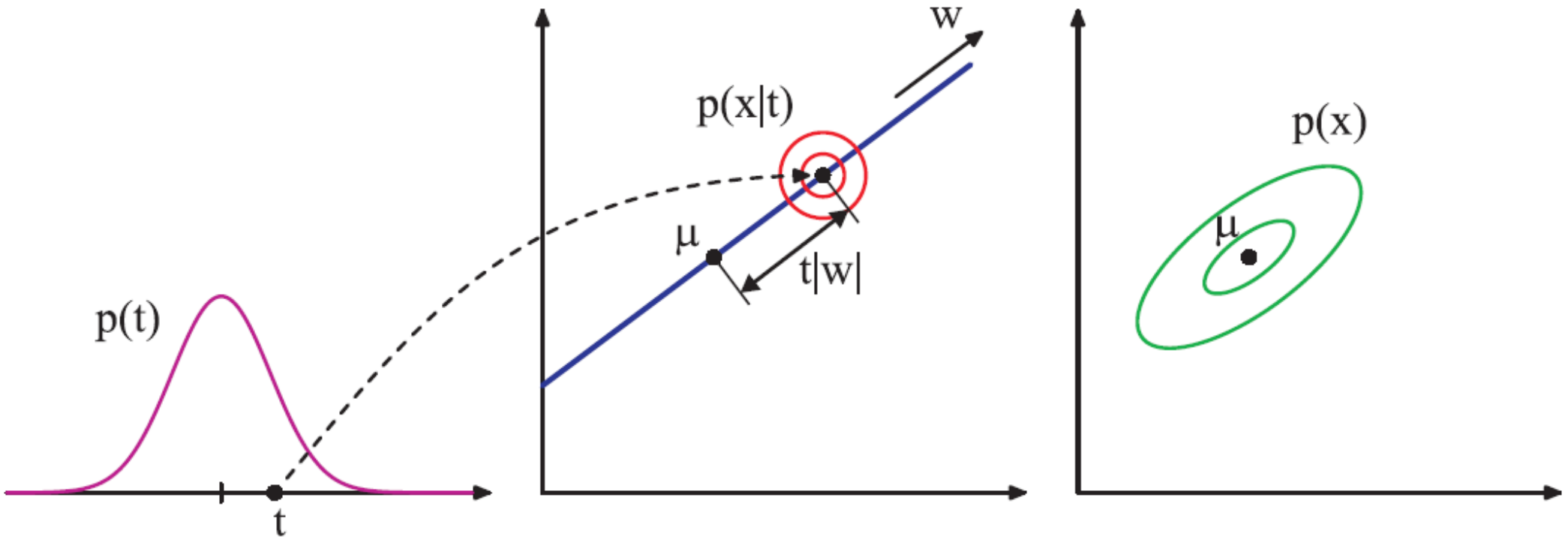
- **Постановка задачі**: Нехай є набір спостережуваних даних $X = \{\boldsymbol{x}_1 \dots \boldsymbol{x}_m\} \in \mathbb{R}^n$. Дослідимо задачу автоматичного обрання *ефективної* розмірності підпростору для їх представлення, а також дослідимо питання зменшення похибки реконструкції в цьому підпросторі порівняно із класичним алгоритмом PCA.

- **Ймовірнісна модель PCA (PPCA)**:
  - набір прихованих змінних $T = \{\boldsymbol{t}_1 \dots \boldsymbol{t}_m\} \in \mathbb{R}^d$, $d < n$, d – бажана розмірність;
  - $p(\boldsymbol{x}|\boldsymbol{t}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{W}\boldsymbol{t} + \boldsymbol{\mu}, \sigma^2\boldsymbol{I}), \qquad p(\boldsymbol{t}) = \mathcal{N}(\boldsymbol{t}|\boldsymbol{0}, \boldsymbol{I}) \quad \Longrightarrow$
    $\Longrightarrow \quad p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{C} = \boldsymbol{W}\boldsymbol{W}^T + \sigma^2\boldsymbol{I} \in \mathbb{R}^{n\times n})$
  - параметри моделі $\boldsymbol{W}, \boldsymbol{\mu}, \sigma$ можна знайти методом максимальної правдоподібності (аналітично, або чисельно за допомогою EM-алгоритму);
  - знайшовши параметри, представлення для $\boldsymbol{x}$ є мат сподіванням $p(\boldsymbol{t}|\boldsymbol{x})$:

$$\mathbb{E}_{\boldsymbol{t}|\boldsymbol{x}}\boldsymbol{t} = \boldsymbol{M}^{-1}\boldsymbol{W}^T(\boldsymbol{x} - \boldsymbol{\mu}), \qquad \boldsymbol{M} = \boldsymbol{W}^T\boldsymbol{W} + \sigma^2\boldsymbol{I} \in \mathbb{R}^{d\times d}$$

Задача 3. Модель PPCA

Генерація об’єктів у моделі PPCA для $n = 2$ для $d = 1$:



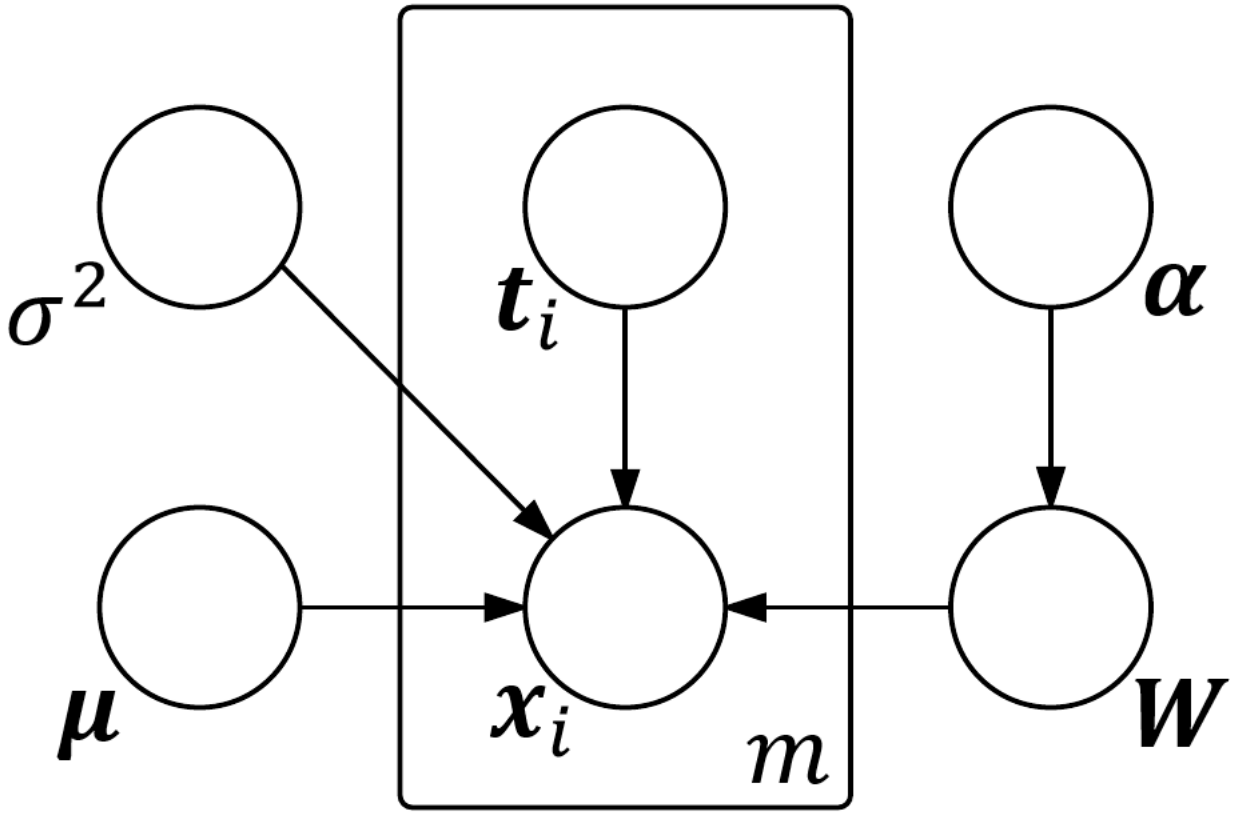
Задача 3. Модель BPCA

Байєсівська модель PCA (BPCA):

- $d := n - 1$
- $p(\boldsymbol{W}|\boldsymbol{\alpha}) = \prod_{i=1}^{n-1} \left(\frac{\alpha_i}{2\pi}\right)^{n/2} \exp\left\{-\frac{1}{2}\alpha_i \|\boldsymbol{w}_i\|^2\right\}$, гіперпараметри α_i контролюють обернені дисперсії відповідних $\boldsymbol{w}_i$: $\alpha_i \longrightarrow +\infty \implies \boldsymbol{w}_i \longrightarrow \boldsymbol{0}$ – можна відкинути
- Параметри знаходимо методом максимальної правдоподібності за допомогою EM-алгоритму:

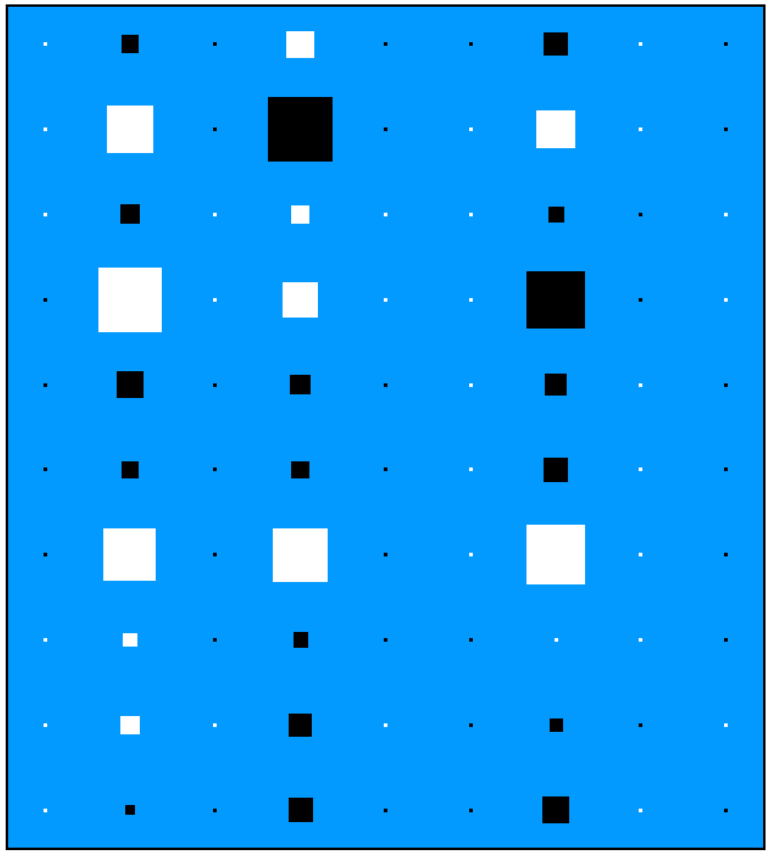
$$p(X|\boldsymbol{\mu}, \sigma, \boldsymbol{\alpha}) = \int p(X|\boldsymbol{W}, \boldsymbol{\mu}, \sigma) p(\boldsymbol{W}|\boldsymbol{\alpha})\, d\boldsymbol{W} \longrightarrow \max_{\boldsymbol{\mu}, \sigma, \boldsymbol{\alpha}}$$

Задача 3. Графічне представлення моделі BPCA



Задача 3. Опис даних та результати

Вхідні дані: $x_1 \dots x_{300} \sim N(\mathbf{0}, \mathrm{diag}\{\sigma_1^2 \dots \sigma_{10}^2\})$, $\sigma_i = 1, i \in \{1,2,3\}$ та $\sigma_i = 0.5$ для решти i



Матриця $\boldsymbol{W}$, обчислена за методом BPCA

Алгоритм	MRE
PCA, $d = 5$	1.2812
PCA, $d = 6$	0.9654
PCA, $d = 7$	0.7196
PCA, $d = 8$	0.5593
PCA, $d = 9$	0.2423
BPCA ($d_{eff} = 3$)	0.1851

Порівняння BPCA із PCA за *середньою похибкою реконструкції* (MRE) – $\frac{1}{m}\sum_{i=1}^{m}\|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\|^2$

Висновки та подальші дослідження

Висновки:

- На прикладі лінійної регресії досліджена залежність якості оцінок за допомогою байєсівського аналізу від ступеня точності апріорних знань
- Для задачі класифікації текстів порівняні алгоритми NBC, MEC із класичним SVM
- Автоматична ідентифікація розмірності підпростору в задачі зменшення даних, зменшення похибки реконструкції порівняно із PCA

Подальші дослідження:

- Для розглянутих задач більш ефективні алгоритми чи їх модифікації
- Розширити коло задач (нейронні мережі)

Дякую за увагу!