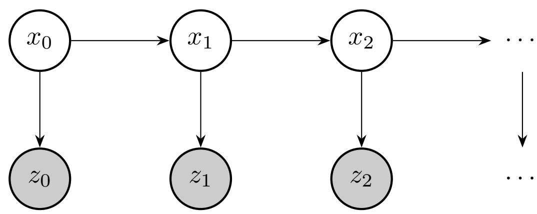
Домашнее задание №2

Задание 1

Количество баллов: 10 (+ *3)



Дана графическая вероятностная модель, в которой

$$\begin{aligned} \mathbf{x}_t &= \mathbf{A}\mathbf{x}_{t-1} + \mathbf{w}_t, & \mathbf{w}_t &\sim \mathcal{N}(\mathbf{0}, \mathbf{Q}) \\ \mathbf{z}_t &= \mathbf{H}\mathbf{x}_t + \mathbf{v}_t, & \mathbf{v}_t &\sim \mathcal{N}(\mathbf{0}, \mathbf{R}) \end{aligned} \tag{1}$$

Для оценки скрытых состояний $\mathbf{x}_t$ по наблюдениям $\mathbf{z}_t$ используется рекуррентный алгоритм фильтрации, известный как *фильтр Калмана.*

1. Prediction step

Пусть дано распределение состояния

$$\mathbf{x}_{t-1} \,|\, \mathbf{z}_{1:t-1} \sim \mathcal{N}(\hat{\mathbf{x}}_{t-1|t-1}, \mathbf{P}_{t-1|t-1}),$$

где $\hat{\mathbf{x}}_{i|j}, \mathbf{P}_{i|j}$ обозначают математическое ожидание и ковариацию состояния $\mathbf{x}_i$ после наблюдения данных $\mathbf{z}_{1:j} = (\mathbf{z}_1, \ldots, \mathbf{z}_j)$.

1. Выпишите распределение $p(\mathbf{x}_t \,|\, \mathbf{z}_{1:t-1})$ в явном виде.
2. Найдите его математическое ожидание $\hat{\mathbf{x}}_{t|t-1}$ и ковариацию $\mathbf{P}_{t|t-1}$ в терминах $\mathbf{A}, \mathbf{Q}$, $\hat{\mathbf{x}}_{t-1|t-1}$ и $\mathbf{P}_{t-1|t-1}$.

2. Update step

1. Выпишите распределение $p(\mathbf{x}_t \,|\, \mathbf{z}_{1:t})$ в явном виде.
2. Найдите математическое ожидание $\hat{\mathbf{x}}_{t|t}$ и ковариацию $\mathbf{P}_{t|t}$ этого распределения.
3. Выпишите в явном виде $\mathbf{K}_t$, т.ч. $\hat{\mathbf{x}}_{t|t} = \hat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t(\mathbf{z}_t - \mathbf{H}\hat{\mathbf{x}}_{t|t-1})$ и $\mathbf{P}_{t|t} = (\mathbf{I} - \mathbf{K}_t\mathbf{H})\mathbf{P}_{t|t-1}$.

3. Implementation

1. Напишите класс в Python (в Jupyter Notebook), реализующий prediction / update step (за основу можно взять `kalman.py`).

2. Эксперимент с синтетикой:

    - Выберите $\mathbf{A}, \mathbf{H}, \mathbf{Q}, \mathbf{R}$, сгенерируйте на их основе траекторию $\{(\mathbf{x}_t, \mathbf{z}_t)\}_{t=1}^{T}$.
    - Вычислите разницу $\left\|\mathbf{x}_t - \hat{\mathbf{x}}_{t|t}\right\|$, проанализируйте её поведение с течением времени, нарисуйте график.
    - Проанализируйте поведение $\mathbf{P}_{t|t}$ с течением времени.

3. Эксперимент с реальными данными:

    - Загрузите данные из файла `data.csv`, содержащего измерения акселерометра и гироскопа.
    - Определите модель состояния и наблюдения:

$$\mathbf{x}_t = \begin{bmatrix} \text{position}_x(t) \\ \text{position}_y(t) \\ \text{position}_z(t) \\ \text{velocity}_x(t) \\ \text{velocity}_y(t) \\ \text{velocity}_z(t) \end{bmatrix}, \quad \mathbf{z}_t = \begin{bmatrix} \text{acceleration}_x(t) \\ \text{acceleration}_y(t) \\ \text{acceleration}_z(t) \end{bmatrix},$$

      и задайте матрицы $\mathbf{A}, \mathbf{H}, \mathbf{Q}, \mathbf{R}$.

    - Примените фильтр Калмана к загруженным данным. Визуализируйте результаты (например, траекторию положения и скорость).
    - ***Дополнительно**
      Исследуйте альтернативные параметризации:
        - Добавьте угловые скорости с гироскопа в вектор состояния $\mathbf{x}_t$
        - Измените $\mathbf{A}$ и $\mathbf{H}$ для учёта дополнительных наблюдений
        - Сравните фильтрацию при разных моделях.

Задание 2

Количество баллов: 5 (+ *5)

Рассмотрим вероятностную модель со скрытыми переменными:

$$p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p(\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2\mathbf{I})\mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I}),$$

где $\mathbf{x} \in \mathbb{R}^D, \quad \mathbf{z} \in \mathbb{R}^d, \quad D \gg d,$ $\mathbf{z}$ играет роль сжатого представления для $\mathbf{x}$. Параметры модели представлены как $\boldsymbol{\theta} = (\mathbf{W}, \boldsymbol{\mu}, \sigma^2)$, где $\mathbf{W} \in \mathbb{R}^{D\times d}, \boldsymbol{\mu} \in \mathbb{R}^D, \sigma^2 > 0$.

Для выполнения задания требуется:

1. Найти распределение $p_{\boldsymbol{\theta}}(\mathbf{x})$

2. Найти апостериорное распределение $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ на E-шаге

3. Вывести формулы для оценки параметров $\mathbf{W}, \boldsymbol{\mu}, \sigma^2$ на M-шаге.

4. ***Дополнительно**

    - Реализовать класс для классического PCA, проанализировать его временную сложность
    - Реализовать класс для предложенного выше EM-алгоритма, проанализировать его временную сложность
    - На выбранном датасете сравнить скорость и качество работы двух алгоритмов

Задание 3 (дополнительное)

Количество баллов: *12

Пусть $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, $\mathbf{x}_n \in \mathbb{R}^D$ – независимая выборка из смеси распределений Стьюдента

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = \sum_{k=1}^{K} w_k \mathcal{T}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \nu), \quad w_k \geq 0, \ \sum_j w_j = 1$$

Рассмотрим следующую вероятностную модель со скрытыми переменными:

$$p_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{T}, \mathbf{Z}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \Big[w_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k/z_n) \mathcal{G}(z_n|\nu/2, \nu/2) \Big]^{t_{nk}}$$

Здесь

- Гиперпараметр ν фиксирован, параметры $\boldsymbol{\theta} = (\mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, где $\mathbf{w} = (w_1, \dots, w_K), \boldsymbol{\mu} = (\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K), \boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_K), \ w_k \geq 0, \ \sum_j w_j = 1$.

- $t_{nk} \in \{0, 1\}$, $\sum_j t_{nj} = 1$ обозначает принадлежность n-го объекта k-ой компоненте смеси.

- $\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k/z_n)$ – плотность нормального распределения с параметрами $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k/z_n)$, вычисленная в точке $\mathbf{x}_n$.

- $\mathcal{G}(z_n|\nu/2, \nu/2)$ – плотность гамма распределения с параметрами $(\nu/2, \nu/2)$, вычисленная в точке z_n.

Можно показать, что

$$p_{\boldsymbol{\theta}}(\mathbf{X}) = \int p_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{T}, \mathbf{Z}) d\mathbf{T} d\mathbf{Z}$$

Поэтому оценку максимального правдоподобия $\boldsymbol{\theta}^{\mathrm{ML}}$ для смеси распределений Стьюдента $p_{\boldsymbol{\theta}}(\mathbf{X})$ можно искать с помощью вариационного EM-алгоритма для модели с латентными переменными $p_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{T}, \mathbf{Z})$, в котором на E-шаге апостериорное распределение приближается в семействе

$$q_{\mathbf{T}}(\mathbf{T}) q_{\mathbf{Z}}(\mathbf{Z}) \approx p_{\boldsymbol{\theta}}(\mathbf{T}, \mathbf{Z} | \mathbf{X}).$$

Для выполнения задания требуется:

1. Выписать формулы пересчёта для компонент вариационного приближения $q_{\mathbf{T}}(\mathbf{T})$ и $q_{\mathbf{Z}}(\mathbf{Z})$

2. Выписать формулы пересчёта параметров $w_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ на M-шаге

3. Выписать функционал $\mathcal{L}(q, \boldsymbol{\theta})$ – нижнюю оценку на $\log p_{\boldsymbol{\theta}}(\mathbf{X})$

4. Найти формулы для статистик распределений $q_{\mathbf{T}}(\mathbf{T})$ и $q_{\mathbf{Z}}(\mathbf{Z})$, требуемых в предыдущих трёх пунктах