Лекция 12

Уменьшение размерности

Машинное обучение

Андрей Фильченков / Сергей Муравьёв

20.11.2020

План лекции

- Уменьшение размерности
- Извлечение: Метод главных компонент
- Извлечение: Автокодировщики
- Извлечение: t-SNE
- Выбор: Встроенные методы
- Выбор: Методы-обертки
- Выбор: Фильтры
- Выбор: Гибриды и ансамбли

- Слайды доступны: **shorturl.at/ltVZ3**
- Видео доступны: **shorturl.at/hjyAX**

План лекции

- Уменьшение размерности
- Извлечение: Метод главных компонент
- Извлечение: Автокодировщики
- Извлечение: t-SNE
- Выбор: Встроенные методы
- Выбор: Методы-обертки
- Выбор: Фильтры
- Выбор: Гибриды и ансамбли

Задача уменьшения размерности

Объекты описаны признаками $\mathcal{F} = (f_1, \dots, f_n)$.

Задача: построить множество признаков $\mathcal{G} = (g_1, \dots, g_k)$: $k < n$ (часто $k \ll n$), при переход к которым сопровождается наименьшей потерей информации.

- Ускорение обучение и обработку
- Борьба с шумом и мультиколлинеарностью
- Интерпретация и визуализация данных

Проклятие размерности

Проклятие размерности (curse of dimensionality) это набор проблем, возникающих с ростом размерности

- Увеличиваются требования к памяти и вычислительной мощности
- Данные становятся более разреженными
- Проще найти гипотезы, не имеющие отношения к реальности

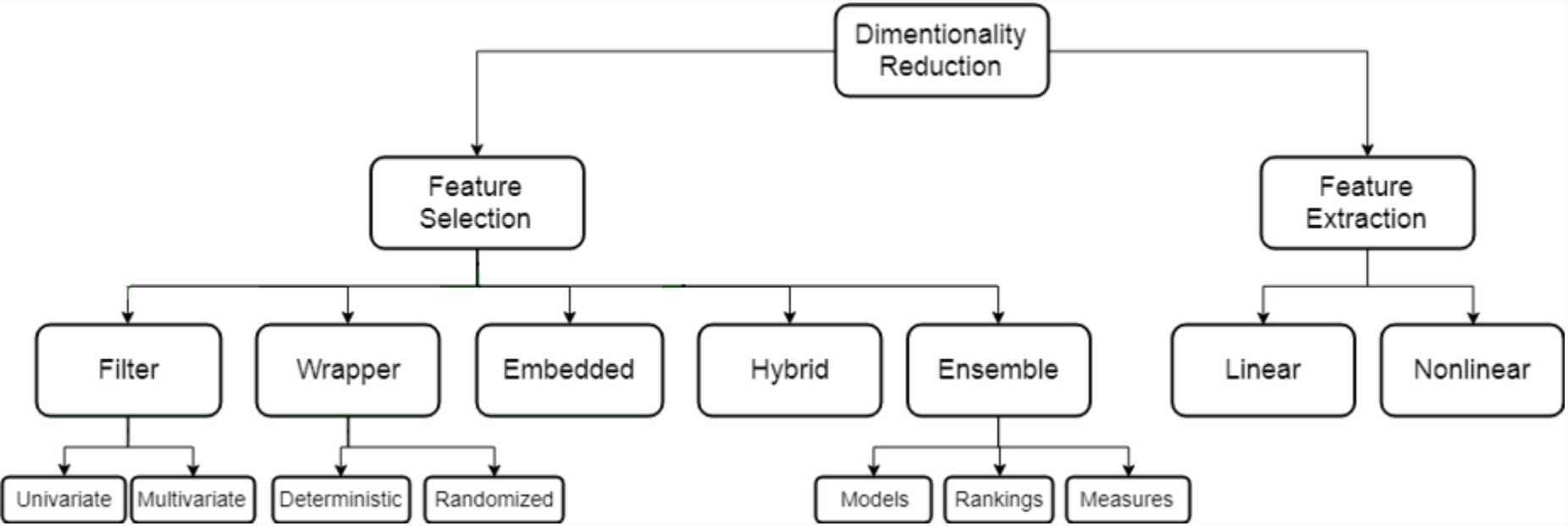
Когда применять

Уменьшение размерности — шаг в предобработки данных

- Меньше памяти для хранения
- Уменьшение времени обработки
- Увеличение качества обработки
- Понимание природы признаков

Обучение представлению это тоже уменьшение размерности, но с учителем

Методы уменьшения размерности



Два основных подхода

Выбор признаков (feature selection) включает методы, для которых $\mathcal{G} \subset \mathcal{F}$. Они

- быстро работают;
- не могут «выдумывать» сложных признаков.

Извлечение признаков (feature extraction) включает все другие методы (в том числе даже те, у которых $k > n$).

- в целом, дольше работают;
- могут извлекать сложные признаки.

Извлечение признаков

Цель методов извлечения признаков:

- Уменьшение числа ресурсов, требуемых для обработки больших данных
- Поиск новых признаков
- Эти признаки могут быть линейными и нелинейными относительно исходных

Выбор признаков

Цели выбора признаков:

- Уменьшение переобучения у улучшение качества предсказания
- Улучшее понимание моделей

Типы ненужных признаков

- **Избыточные (redundant) признаки** не привносят дополнительной информации относительно существующих
- **Нерелевантные (irrelevant) признаки** просто неинформативны

План лекции

- Уменьшение размерности
- Извлечение: Метод главных компонент
- Извлечение: Автокодировщики
- Извлечение: t-SNE
- Выбор: Встроенные методы
- Выбор: Методы-обертки
- Выбор: Фильтры
- Выбор: Гибриды и ансамбли

Мотивация извлечения признаков

Соберем большой набор данных размерности 50

Country	GDP (trillions of US$)	Per capita GDP (thousands of intl. $)	Human Develop-ment Index	Life expectancy	Poverty Index (Gini as percentage)	Mean household income (thousands of US$)	...
Canada	1.577	39.17	0.908	80.7	32.6	67.293	...
China	5.878	7.54	0.687	73	46.9	10.22	...
India	1.632	3.41	0.547	64.7	36.8	0.735	...
Russia	1.48	19.84	0.755	65.5	39.9	0.72	...
Singapore	0.223	56.69	0.866	80	42.5	67.1	...
USA	14.527	46.86	0.91	78.3	40.8	84.3	...
...	...	...	...	...	...	...	

Синтез признаков

Извлечение признаков позволяет получить другое представление объектов

Country	z_1	z_2
Canada	1.6	1.2
China	1.7	0.3
India	1.6	0.2
Russia	1.4	0.5
Singapore	0.5	1.7
USA	2	1.5
...	...	...

Извлечение линейных признаков

Бывает линейным и нелинейным

Линейный быстрее и интерпретируемее

Нелинейные находят более сложные признаки

Метод главных компонент (Principal Component Analysis, PCA) наиболее известный метод для извлечения линейных признаков

Одномерный случай

Дано множество точек в R^n. Хотим описать эти данные при помощи одной переменной.

Это задача регрессии без учителя.

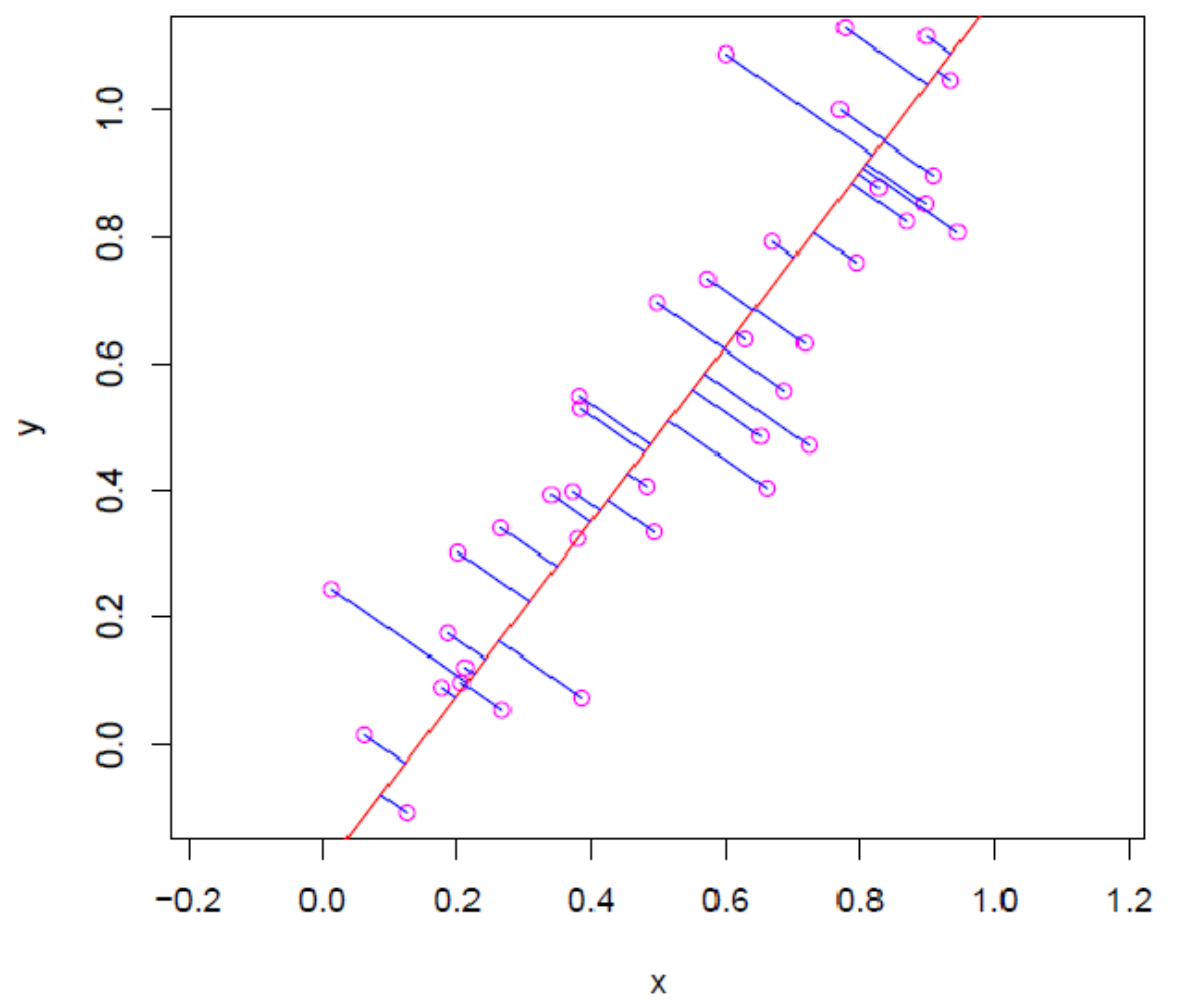
Основная идея: будем проецировать на прямую, такую, что

1) расстояние от точек до нее минимально;

2) дисперсия проекций максимальна.

Эти условия эквивалентны.

Иллюстрация



Общий случай

Приблизить данные линейным многообразием меньшего размера:

- минимизация расстояния
- максимизация дисперсии проекций
- максимизация расстояния между проекциями
- корреляция между осями проекций равна нулю (новинка!)

Много вариантов того, как можно описать требования к решению

Математическая постановка

Постановка:

$$\|GU^{\top} - F\|^2 \to \min_{G,U},$$

где $F = \begin{pmatrix} f_1(x_1) & \dots & f_1(x_{|\mathcal{D}|}) \\ \dots & \dots & \dots \\ f_n(x_1) & \dots & f_n(x_{|\mathcal{D}|}) \end{pmatrix},$

$$G = \begin{pmatrix} g_1(x_1) & \dots & g_1(x_{|\mathcal{D}|}) \\ \dots & \dots & \dots \\ g_k(x_1) & \dots & g_k(x_{|\mathcal{D}|}) \end{pmatrix},$$

$\operatorname{rank}(U) = \operatorname{rank}(G) = k.$

Основная теорема

Теорема:
если $k \leq \text{rank}(F)$, то минимум достигается, когда столбцы U являются собственными столбцами $F^\top F$, соответствующим k максимальным собственным значениям, и $G = UF$.

Следствия:

1. $U^\top U = \mathbf{E}_r$.
2. $G^\top G = \Lambda = \text{diag}(\lambda_1, \dots, \lambda_r)$.
3. $U\Lambda = F^\top FU; G\Lambda = F^\top FG$.
4. $\|GU^\top - F\|^2 = \|F\|^2 - \text{tr}\,\Lambda = \sum_{r+1}^{n} \lambda_i$.

Главные компоненты

G — линейное многообразие
Оси G — **главные компоненты**

Итеративный поиск главных компонент:

найти прямую c_1, расстояние до которой минимально.

Повторять: найти прямую c_i, ортогональную $\{c_j\}_{j=1}^{i-1}$, расстояние до которой минимально.

Выбор k

Проблема подобна выбору k в EM.

Отсортировать собственные значения $F^\top F$ по убыванию: $\lambda_{(1)} \geq \cdots \geq \lambda_{(n)}$.

$$E(k) = \frac{\|F - UG^\top\|^2}{\|F\|^2} = \frac{\lambda_{(k+1)} + \cdots + \lambda_{(n)}}{\lambda_{(1)} + \cdots + \lambda_{(n)}}$$

$E(k)$ характеризует долю информацию, теряемую при проекции.

Значение k можно выбрать по $E(k)$.

Обсуждение PCA

- Линейное преобразование со всеми достоинствами и недостатками
- Работает относительно недолго
- Широко распространено для сжатия данных и визуализации
- Улучшения: нелинейные методы (главные кривые и главные многообразия)

Вариации и родственники PCA

- Анализ независимых компонент (ICA)
- EM PCA
- Ядерный PCA
- Канонический корреляционный анализ (CCA)

План лекции

- Уменьшение размерности
- Извлечение: Метод главных компонент
- Извлечение: Автокодировщики
- Извлечение: t-SNE
- Выбор: Встроенные методы
- Выбор: Методы-обертки
- Выбор: Фильтры
- Выбор: Гибриды и ансамбли

Автокодировщик

Автокодировщик (autoencoder) — глубокая нейронная сеть, способная строить низкоразмерные представления данных за счет нелинейной трансформации.

Основная идея: заставим сеть предсказывать (восстанавливать) то, что подается ей на вход, ограничив возможность обучиться тривиальному преобразованию.

Ограничение преобразования

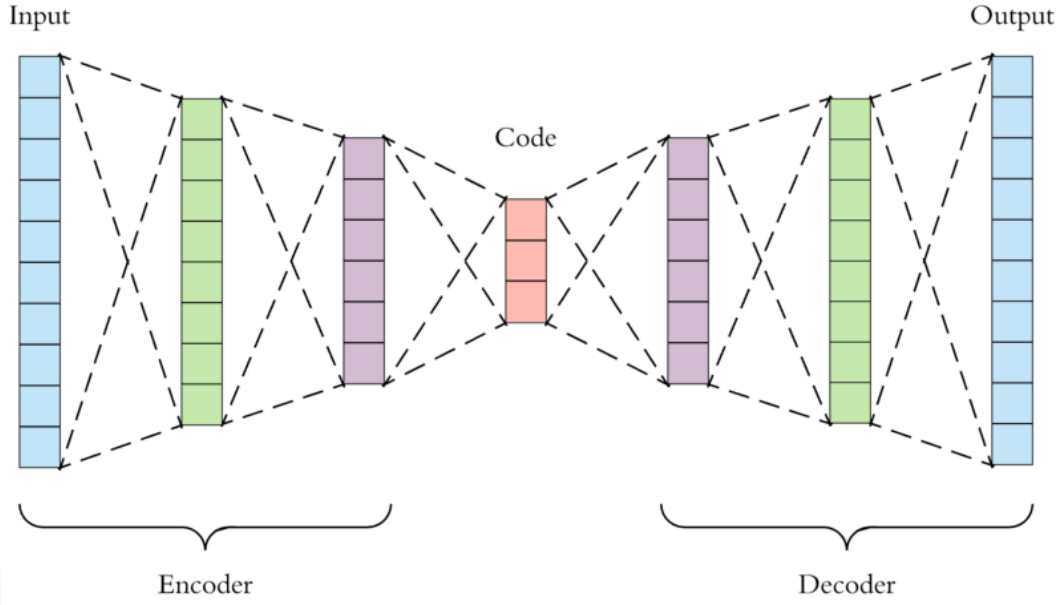
Два варианта:

- Структурный: между входными и выходными слоями должен быть слой меньшей размерности, т.н. **бутылочное горлышко (bottleneck).**
  Это **недополненный (undercomplete)** автокодировщик.
- Регуляризационный: добавим регуляризационную константу к выходам этого слоя, уменьшающим его размерность.
  Это **разреженный (sparse) автокодировщик.**

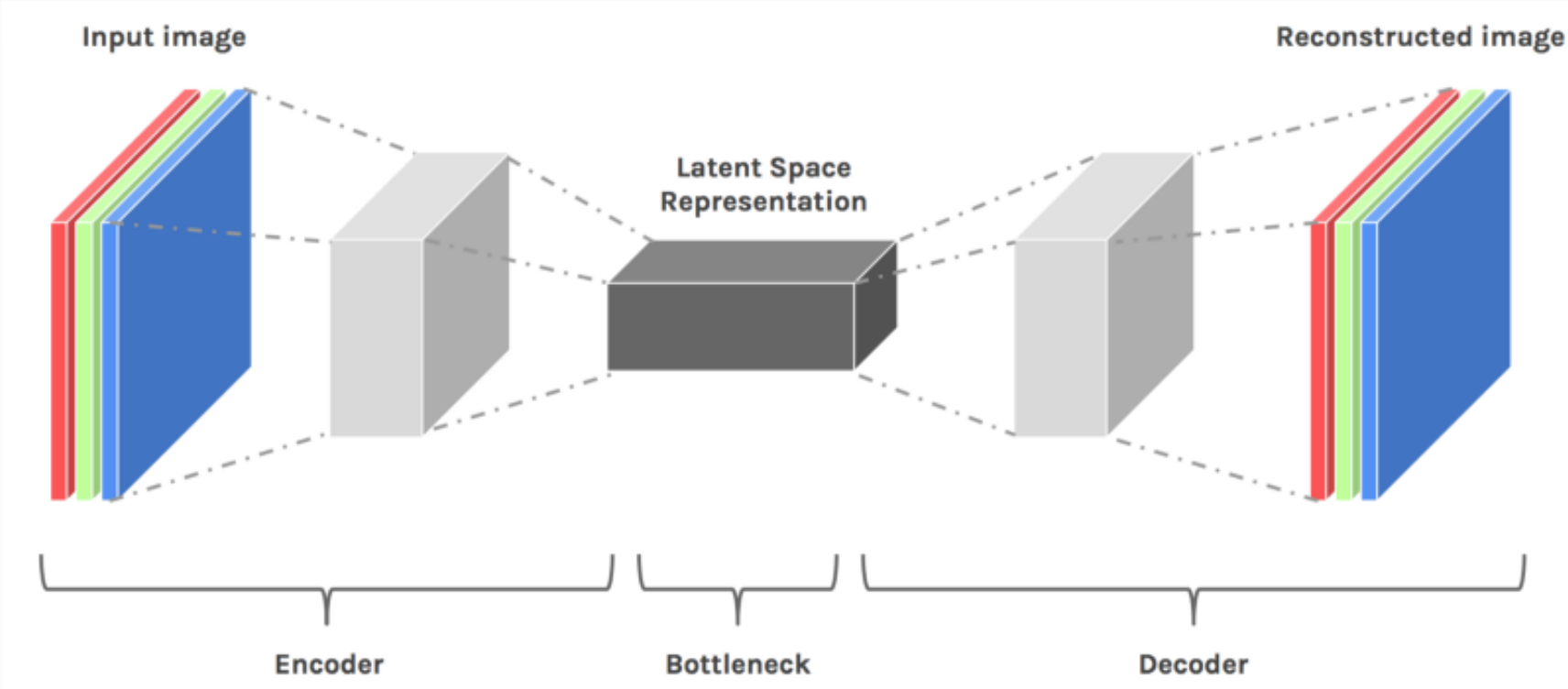
Части автокодировщика

Кодировщик (encoder) — часть сети от входного слоя до бутылочного горлышка

Декодировщик (decoder) — часть сети от бутылочного горлышка до выходного слоя



Модель свёрточного кодировщика



Регуляризация для автокодировщика

Вместо минимизации $\|d(c(x)) - x\|$ будем минимизировать

$$\|d(c(x)) - x\| + \tau \cdot L(c(x)),$$

где c — кодировщик, d — декодировщик, L — некая регуляризация, τ — коэффициент регуляризации.

Стандартно можно взять L_1 норму (как в LASSO)

Вариации автокодировщика

- Шумоподавляющий (denoising) автокодировщик
- Сжимающий (contractive) автокодировщик
- Вариационный (variational) автокодировщик (основан на совсем других принципах!)

План лекции

- Уменьшение размерности
- Извлечение: Метод главных компонент
- Извлечение: Автокодировщики
- Извлечение: t-SNE
- Выбор: Встроенные методы
- Выбор: Методы-обертки
- Выбор: Фильтры
- Выбор: Гибриды и ансамбли

t-SNE

Стохастическое вложение соседей с t-распределением (t-distributed stochastic neighbor embedding, t-SNE) — это алгоритм уменьшения размерности

- Нелинейный
- Используется для визуализации
- Пытается сохранять метрические отношения между объектами

Идея t-SNE

1. Определим вероятность для точки «выбрать ближайшим соседом» другую точку в пространстве
2. Построим такие распределения для высокоразменых и низкоразмерных представлений
3. Минимизируем расстояние между двумя распределениями

Расстояние Кульбака — Лейблера

Расстояние (дивергенция) Кульбака — Лейблера (KL divergence) — расстояние между двумя распределениями P и Q:

$$D_{\mathrm{KL}}(P||Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} \mathrm{d}(x),$$

где p распределено согласно P, а q — согласно Q.

Также называется **относительной энтропией**.

Стохастическое вложение соседей

Определим распределения для обоих пространств так:

$$p_{j|i} = \frac{\exp\left(-\left\|x_i - x_j\right\|^2 / 2\sigma^2\right)}{\sum_{k \neq j} \exp(-\|x_i - x_k\|^2 / 2\sigma^2)}$$

$$q_{j|i} = \frac{\exp\left(-\left\|y_i - y_j\right\|^2\right)}{\sum_{k \neq j} \exp(-\|y_i - y_k\|^2)}$$

Симметричное стохастическое вложение соседей

Определим распределения для обоих пространств так:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2|X|}$$

$$q_{ij} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k \neq l} \exp(-\|y_k - y_l\|^2)}$$

Симметричное стохастическое вложение соседей

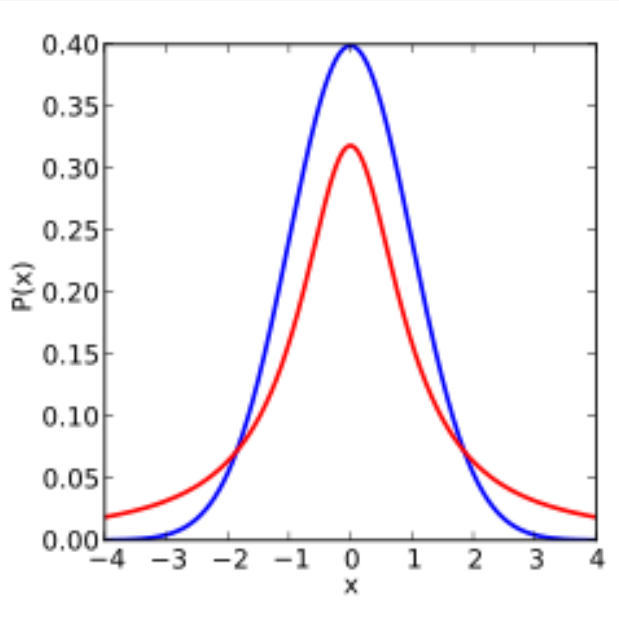
Определим распределения для обоих пространств так:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2|X|}$$

$$q_{ij} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k \neq l} \exp(-\|y_k - y_l\|^2)}$$

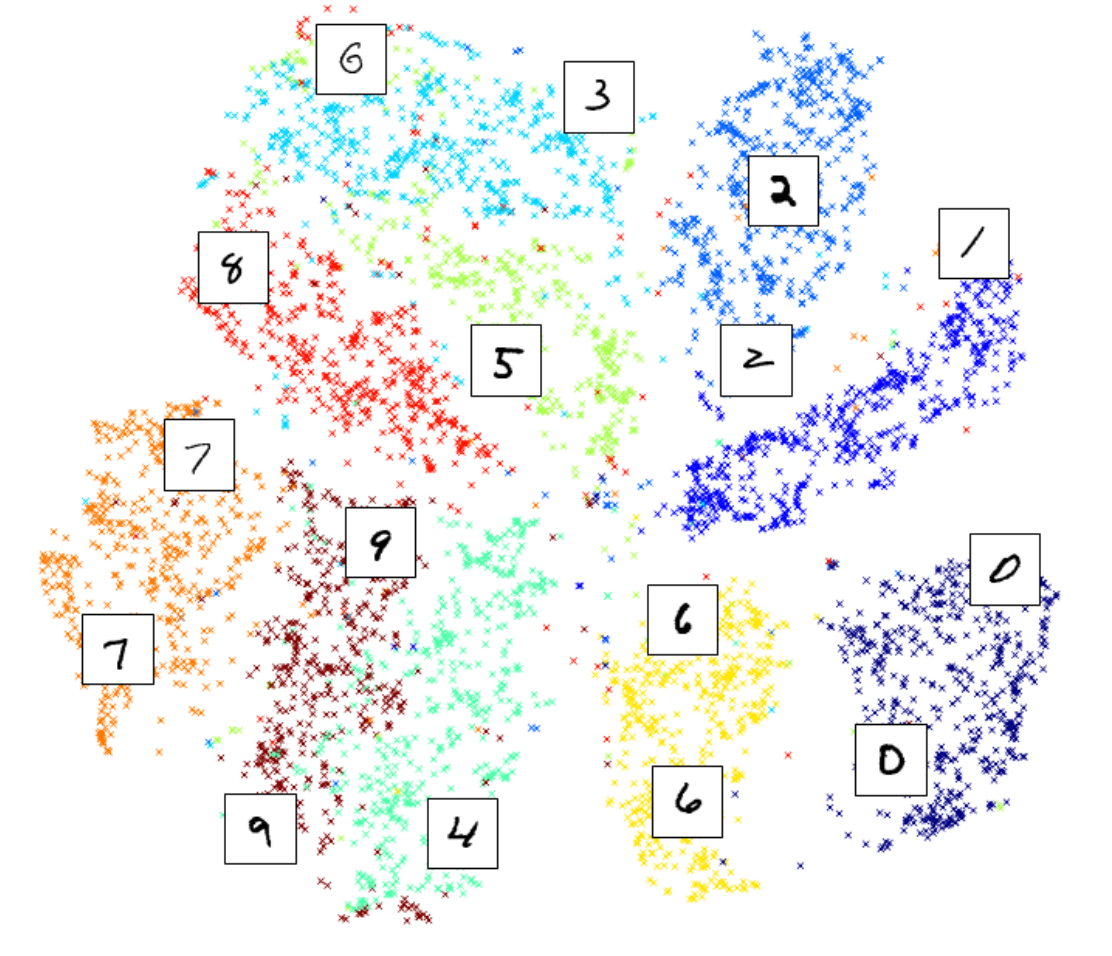
Симметричное стохастическое вложение соседей с t-распределением

Заменим распределение на t-распределение Стьюдента



$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq l}(1 + \|y_k - y_l\|^2)^{-1}}$$

t-SNE на MNIST



План лекции

- Уменьшение размерности
- Извлечение: Метод главных компонент
- Извлечение: Автокодировщики
- Извлечение: t-SNE
- Выбор: Встроенные методы
- Выбор: Методы-обертки
- Выбор: Фильтры
- Выбор: Гибриды и ансамбли

Классификация методов выбора признаков

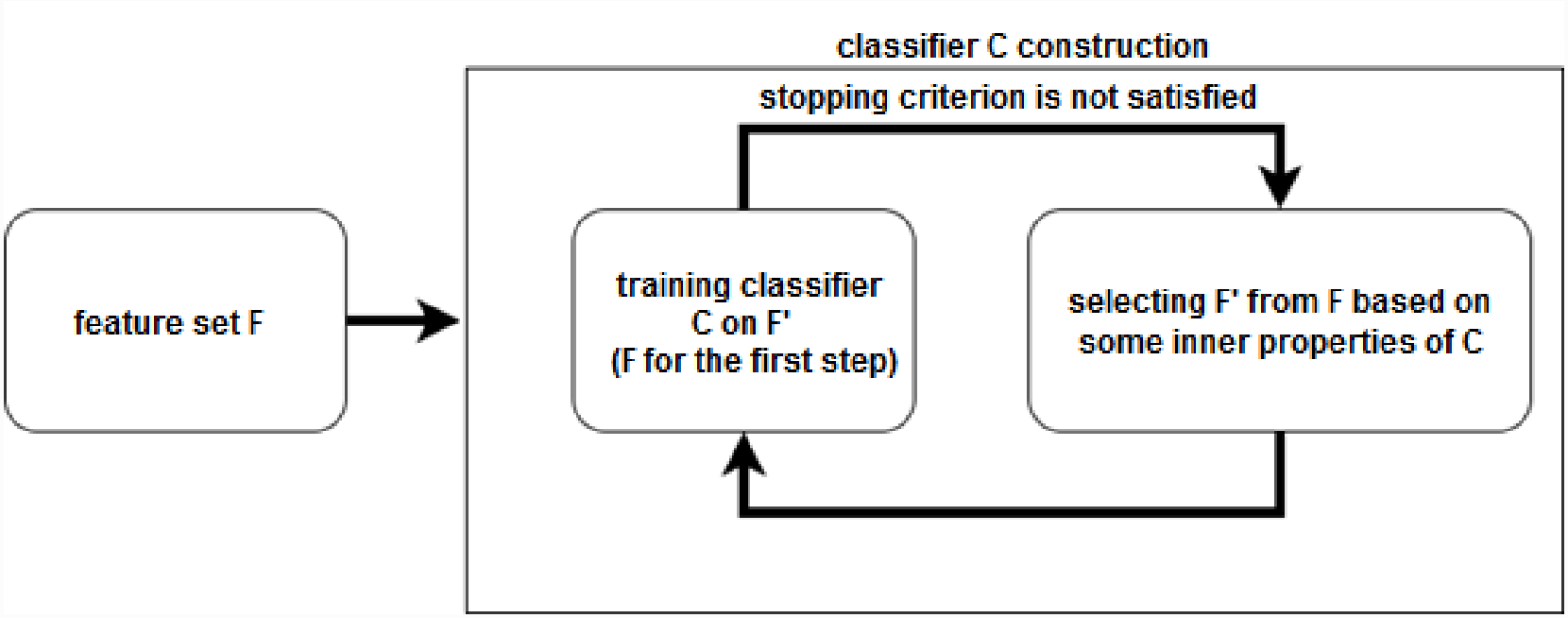
- Встроенные методы (embedded)
- Фильтрующие методы (filter)
    a. Одномерные (univariate)
    b. Многомерные (multivariate)
- Методы-обертки (wrapper)
    a. Детерминированные (deterministic)
    b. Стохастические (stochastic)
- Гибридные и ансамблирующие методы

Встроенные методы

Встроенные методы (embedded methods) это методы выбора признаков, при которых этот выбор осуществляется в процессе работы других алгоритмов (классификаторов и регрессоров)

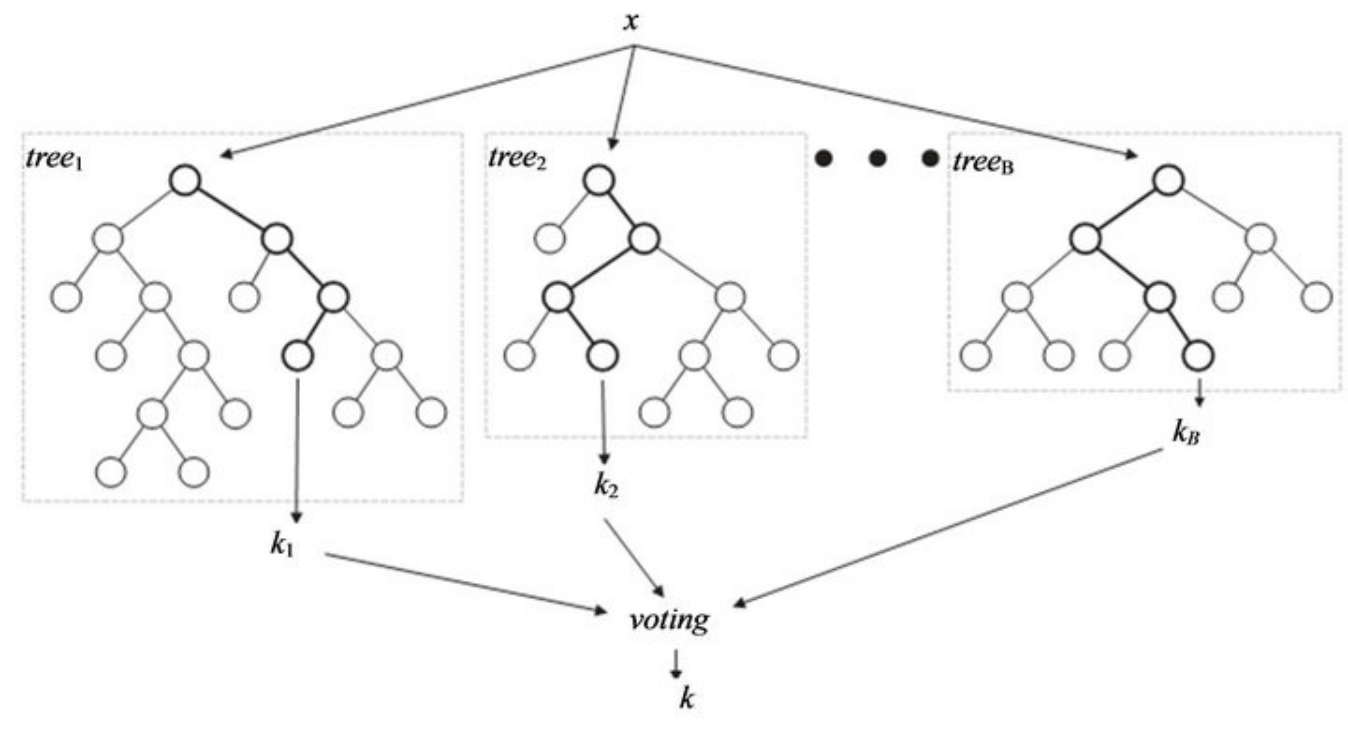
- Опираются на конкретный алгоритм
- Специфичны для каждого алгоритма

Схема встроенного метода



Пример: случайный лес

Каждое дерево выбирает поднабор признаков. Лес также выбирает поднабор



Пример: SVM-RFE

- Обучить SVM на обучающем подмножестве
- Отранжировать признаки согласно их весам
- Выбросить некоторое число признаков с наименьшими весами
- Повторять, пока не останется нужное число признаков

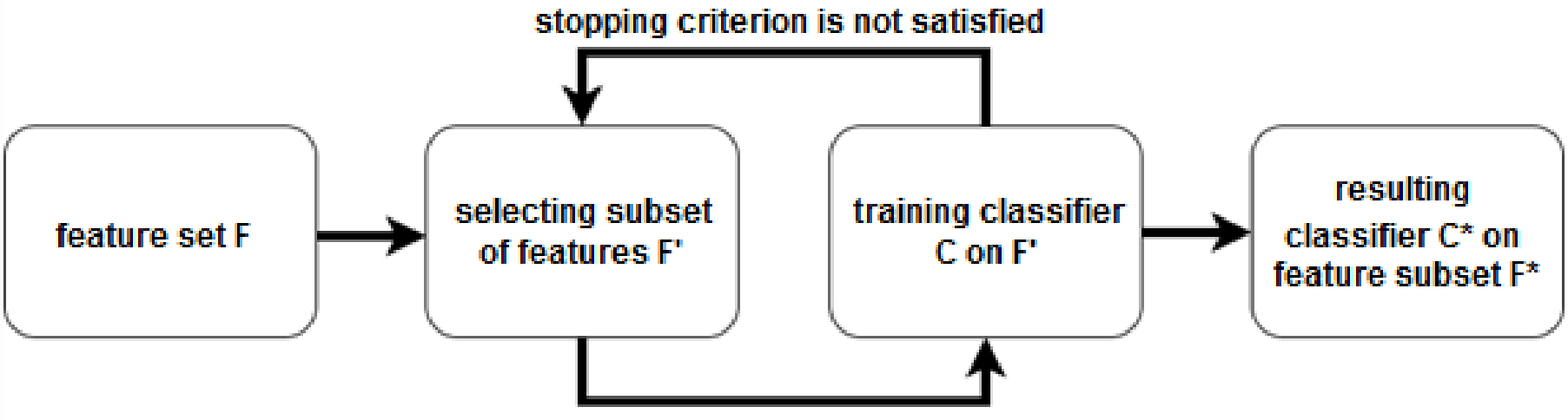
План лекции

- Уменьшение размерности
- Извлечение: Метод главных компонент
- Извлечение: Автокодировщики
- Извлечение: t-SNE
- Выбор: Встроенные методы
- Выбор: Методы-обертки
- Выбор: Фильтры
- Выбор: Гибриды и ансамбли

Метод-обертка

Метод-обертка (wrapper method) использует алгоритм (классификатор или регрессор) для оценки качества получаемого подмножества признаков и использует алгоритмы дискретной оптимизации для поиска оптимального подмножества признаков.

Схема метода-обертки



Классификация методов-оберток

- Детерминированные:
  - SFS (sequential forward selection)
  - SBE (sequential backward elimination)
  - SVM-RFE
- Стохастические:
  - Стохастический поиск восхождением на холм (stochastic hill climbing)
  - Генетические алгоритмы

Анализ методов-оберток

Достоинства:

- Более высокая точность, чем у фильтров
- Используют отношения между признаками
- Оптимизируют качество предсказательной модели в явном виде

Недостатки:

- Очень долго работают
- Могут переобучиться при неправильной работе с разбиением набора данных

План лекции

- Уменьшение размерности
- Извлечение: Метод главных компонент
- Извлечение: Автокодировщики
- Извлечение: t-SNE
- Выбор: Встроенные методы
- Выбор: Методы-обертки
- Выбор: Фильтры
- Выбор: Гибриды и ансамбли

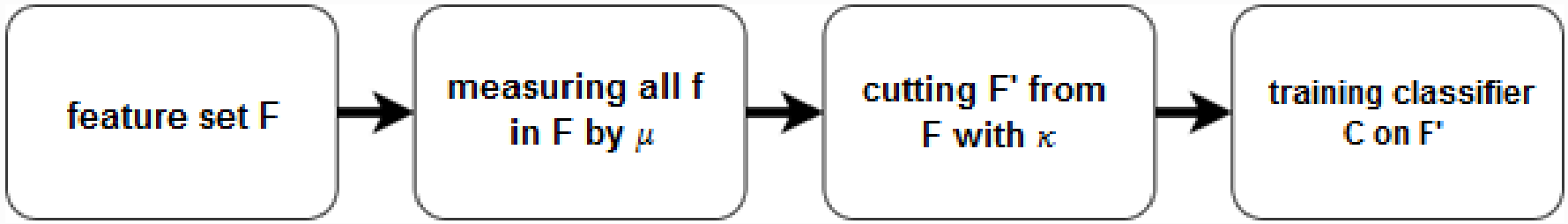
Методы фильтрации

Фильтры (filter methods) оценивают качество отдельных признаков или подмножеств признаков и удаляют худшие

Две компоненты:

- мера значимости признаков μ
- правило обрезки κ определяет какие признаки удалить на основе μ

Схема фильтрующих методов



Классификация фильтрующих методов

- Одномерные (univariate):
  - Евклидово расстояние
  - Прирост информации (IG)
  - Коэффициент корреляции Спирмана
- Многомерные (multivariate):
  - Выбор признаков на основе корреляций (CFS)
  - Фильтр марковского одеяла (MBF)

Корреляция Спирмана

Коэффициент корреляции Спирмана

$$\rho = \frac{\sum_{ij}(x_{ij}-\bar{x}_j)(y_i-\bar{y})}{\sqrt{\sum_{ij}(x_{ij}-\bar{x}_j)^2 \sum_i (y_i-\bar{y})^2}}$$

$$\rho \in [-1; 1]$$

$$\rho \to 0$$

Правило обрезки *к*

Может быть любым

В большинстве случаев используется:

- Число признаков
- Порог значимости признаков

Анализ одномерных фильтров

Преимущества:

- Исключительно быстро работают
- Позволяют оценивать значимость каждого признака

Недостатки:

- Игнорируют отношения между признаками и то, что реально использует предсказательная модель

Анализ многомерных фильтров

Преимущества:

- Работают достаточно быстро
- Учитывают отношения между признаками

Недостатки:

- Работают существенно дольше фильтров
- Не учитывают то, что реально использует предсказательная модель

План лекции

- Уменьшение размерности
- Извлечение: Метод главных компонент
- Извлечение: Автокодировщики
- Извлечение: t-SNE
- Выбор: Встроенные методы
- Выбор: Методы-обертки
- Выбор: Фильтры
- Выбор: Гибриды и ансамбли

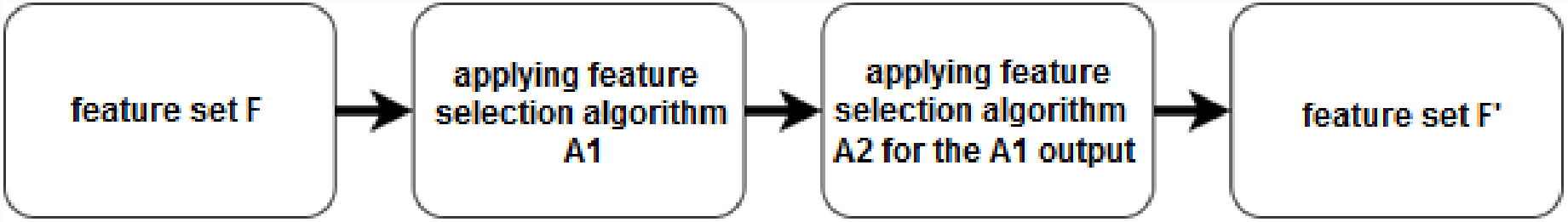
Гибридный подход

Будем комбинировать подходы, чтобы использовать их сильные стороны

Самый частый вариант:

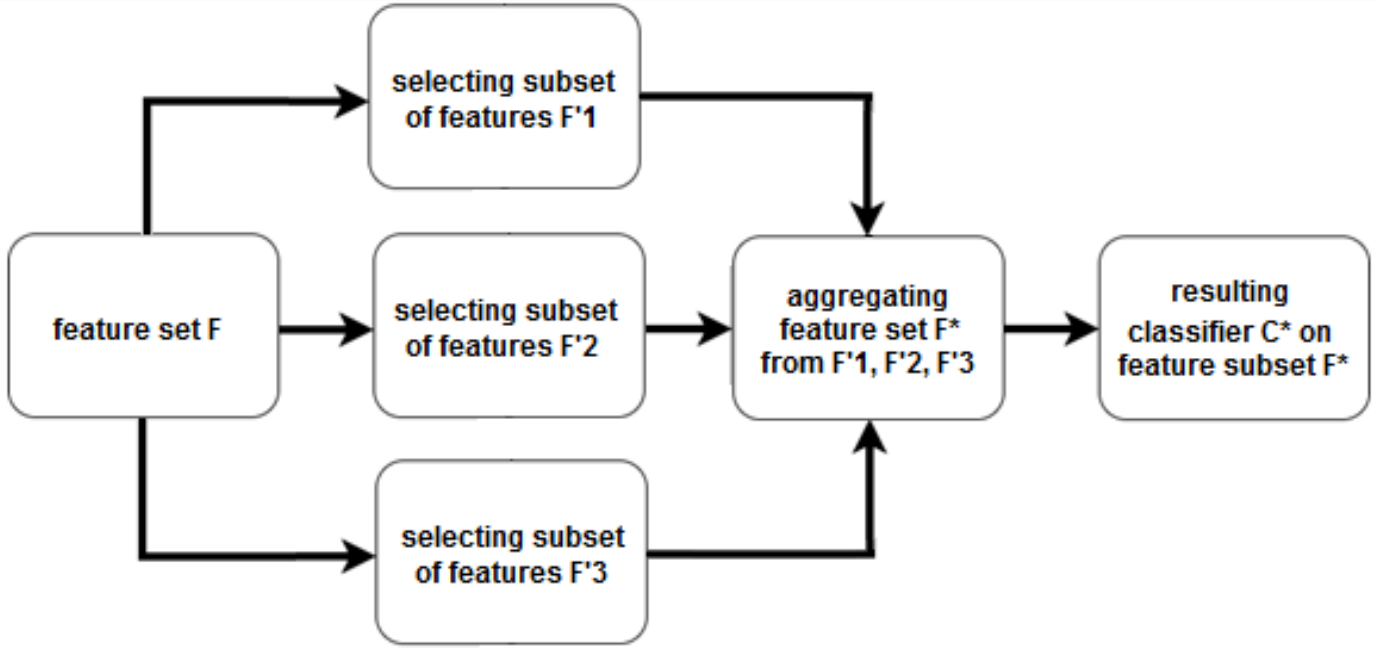
- сначала применим фильтр (или набор фильтров), отсеяв лишние признаки
- затем применим метод-обертку или встроенный метод

Схема гибридного подхода



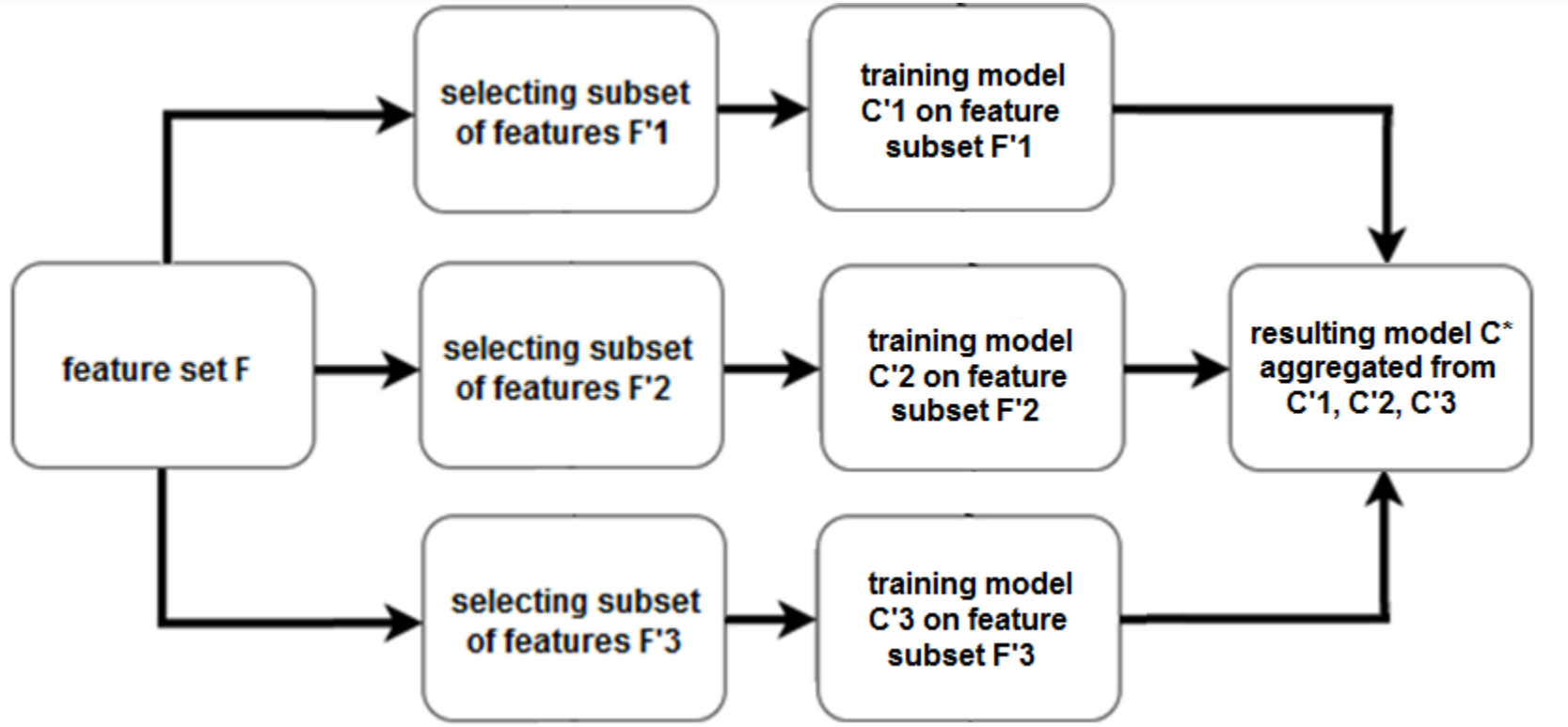
Ансамблирование в выборе признаков

Подход к ансамблированию состоит в построении ансамбля алгоритмов выбора признаков



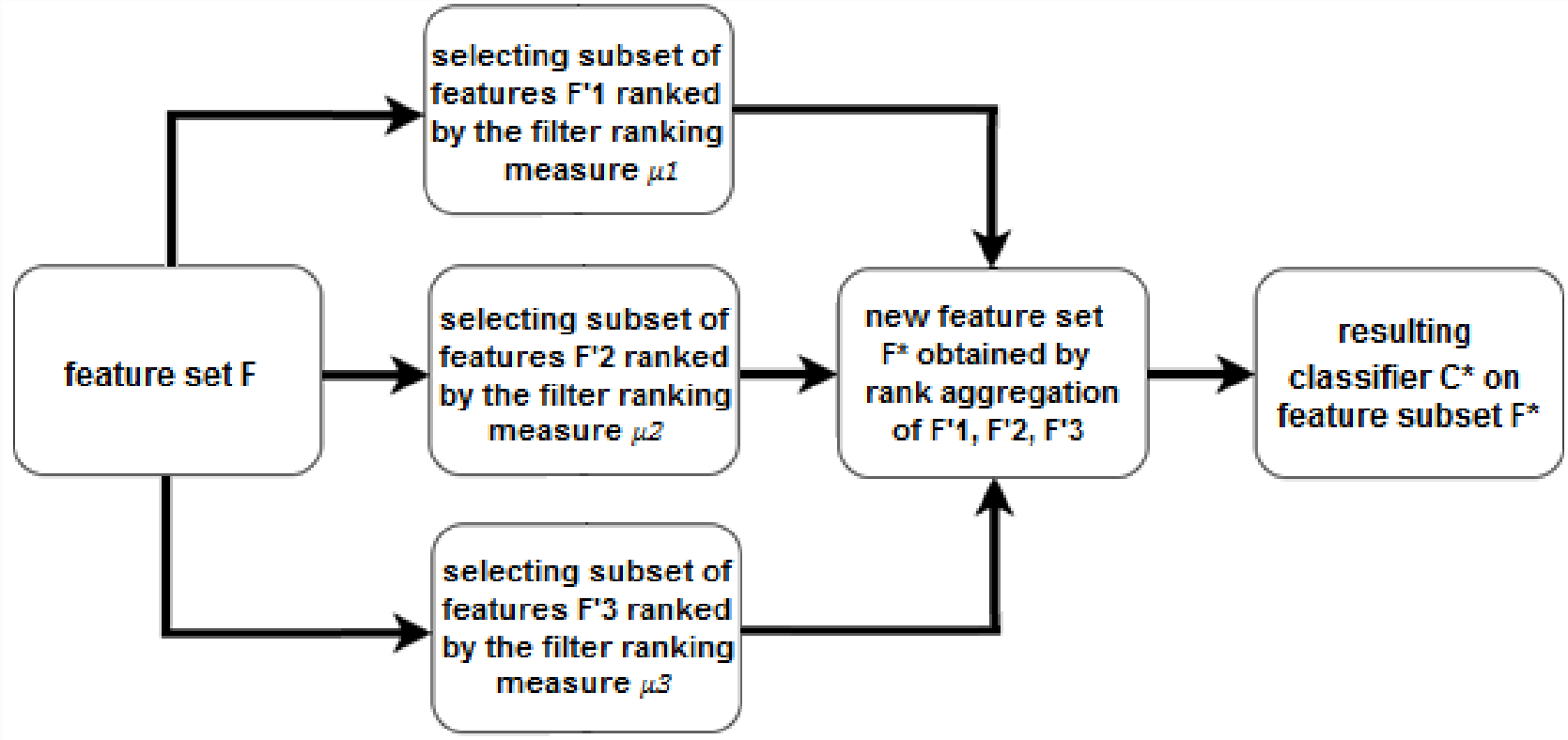
Ансамбль на уровне моделей

Строим ансамблей предсказательных моделей



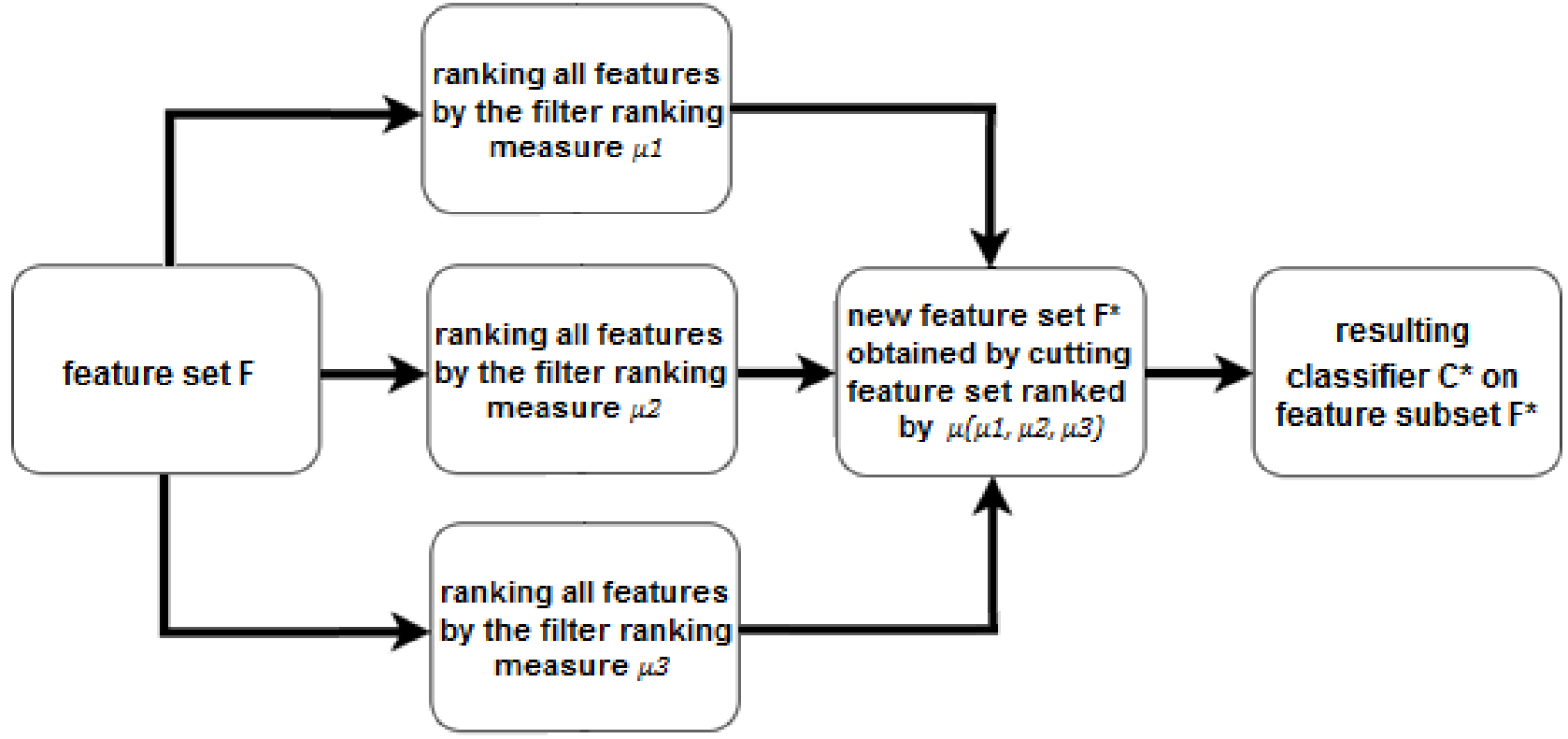
Ансамбль на уровне ранжирований

Объединяем ранжирования



Ансамбль на уровне мер значимости

Объединяем меры значимости



Анализ гибридных и ансамблирующих методов

Преимущества:

- Чаще всего лучше по времени и по качеству

Недостатки:

- Иногда теряется интерпретируемость
- Иногда требуется заботиться о проблеме переобучения