

ЗАЧЕМ ОТБИРАТЬ ПРИЗНАКИ?

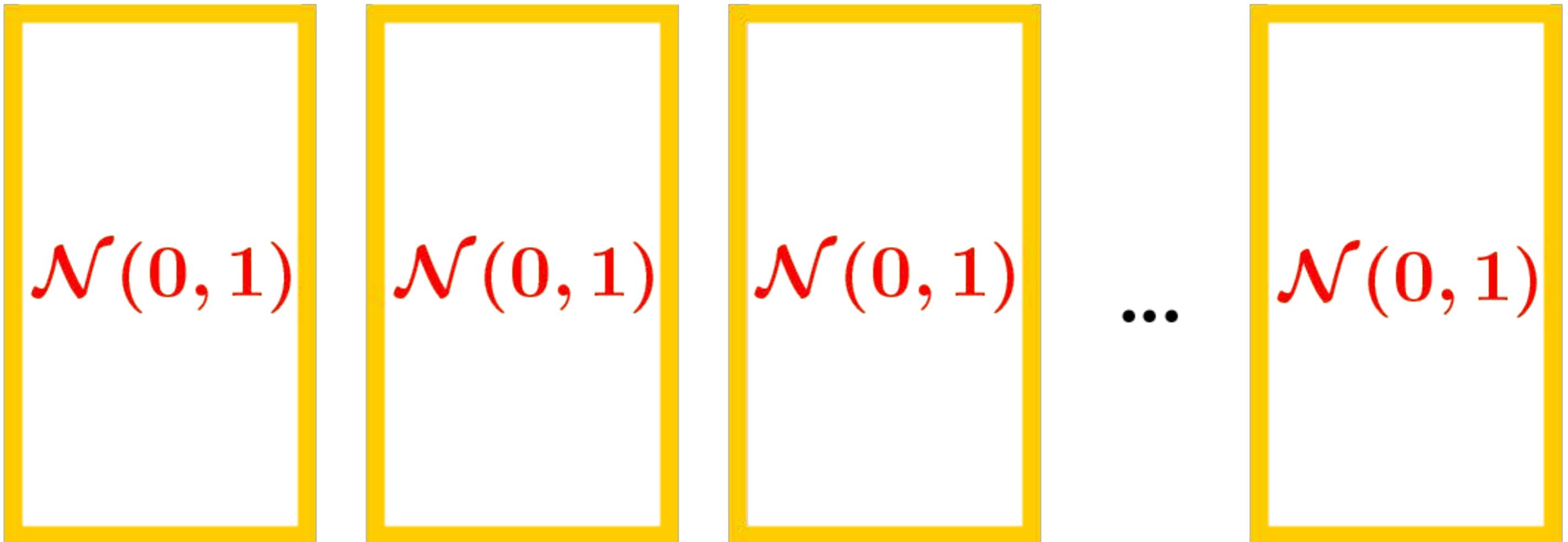
ЧТО ЭТО ТАКОЕ?

- Признаков может быть больше, чем нужно
- Можно повысить качество, если:
 - ▶ Отобрать только полезные
 - ▶ Сформировать новые на основе старых

ШУМОВЫЕ ПРИЗНАКИ

- › Признаки, которые никак не связаны с целевой переменной
- › Но по обучающей выборке это не всегда можно понять

ШУМОВЫЕ ПРИЗНАКИ



РЕШАЮЩЕЕ ДЕРЕВО

- › 1000 признаков
- › Дерево глубины 10 учитёт каждый лишь по одному разу
- › Получатся тысячи листьев
- › В каждый должны попасть много объектов — иначе риск переобучения

УСКОРЕНИЕ МОДЕЛЕЙ

- Чем больше признаков, тем сложнее модели
- Чем сложнее модели, тем дольше они вычисляют прогнозы
- Могут быть жёсткие ограничения на скорость

ОДНОМЕРНЫЕ МЕТОДЫ

- Измерить связь каждого признака с целевой переменной
- Не учитываются сложные закономерности

ОТБОР ПРИЗНАКОВ

➤ Надстройка над обучением:

- ▶ Перебираем комбинации признаков
- ▶ Для каждой обучаем модель
- ▶ Выбираем комбинацию, дающую лучшую модель

ОТБОР ПРИЗНАКОВ

➤ Отбор с помощью модели:

- ▶ L_1 -регуляризация
- ▶ Решающие деревья
- ▶ Случайные леса, градиентный бустинг

ПОНИЖЕНИЕ РАЗМЕРНОСТИ

- Что, если все признаки важны?
- Можно сгенерировать новые на их основе!
- Пример: линейные комбинации

РЕЗЮМЕ

- Понижение размерности — для повышения качества и ускорения моделей
- Отбор признаков — поиск наиболее важных
- Понижение размерности — формирование новых признаков на основе исходных

ОДНОМЕРНЫЙ ОТБОР ПРИЗНАКОВ

ОБОЗНАЧЕНИЯ

- x_{ij} — значение j -го признака на i -м объекте
- \bar{x}_j — среднее значение j -го признака
- y_i — значение целевой переменной на i -м объекте
- \bar{y} — среднее значение целевой переменной

ОДНОМЕРНЫЙ ОТБОР

- Оценить предсказательную силу (информативность) каждого признака
- Затем:
 - ▶ Отобрать k лучших признаков
 - ▶ Отобрать признаки, у которых «сила» выше порога

КОРРЕЛЯЦИЯ

$$R_j = \frac{\sum_{i=1}^{\ell} (x_{ij} - \bar{x}_j)(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{\ell} (x_{ij} - \bar{x}_j)^2 \sum_{i=1}^{\ell} (y_i - \bar{y})^2}}$$

- » Чем больше $|R_j|$, тем информативнее признак
- » Учитывает только линейную связь

КОРРЕЛЯЦИЯ

- › Для вещественных признаков и ответов
- › Для бинарных признаков или ответов — можно использовать значения $\{-1, +1\}$

БИНАРНАЯ КЛАССИФИКАЦИЯ

- Строим классификатор над одним признаком
- Измеряем качество
- Классификатор: $a(\mathbf{x}_i) = [\mathbf{x}_{ij} < t]$
- Качество: AUC-ROC

ВЗАИМНАЯ ИНФОРМАЦИЯ

- › Значения признака: $1, 2, \dots, n$
- › Значения целевой переменной: $1, 2, \dots, m$

$$P(x = v, y = k) = \frac{1}{\ell} \sum_{i=1}^{\ell} [x_{ij} = v][y_i = k]$$

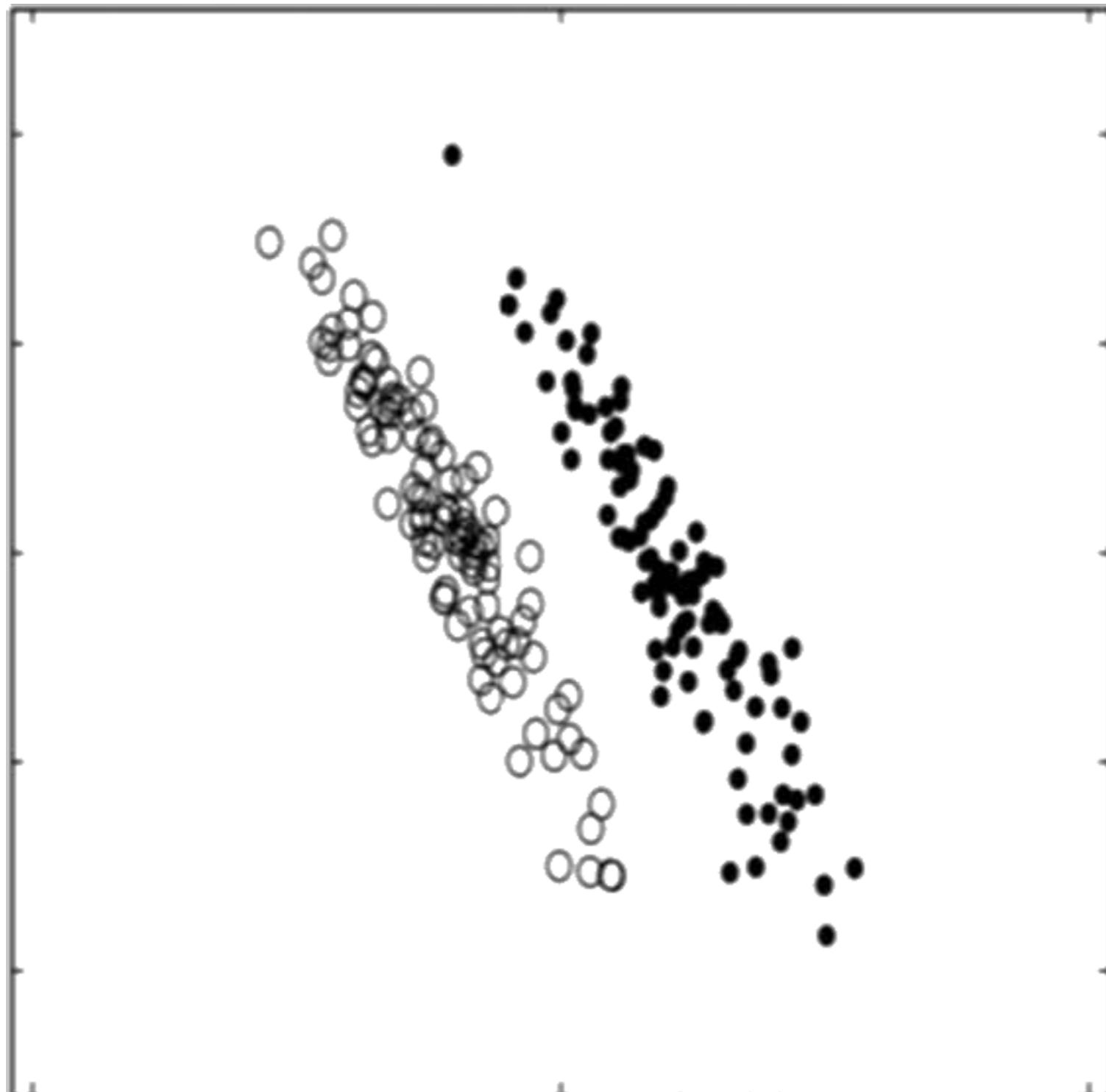
$$P(x = v) = \frac{1}{\ell} \sum_{i=1}^{\ell} [x_{ij} = v]$$

ВЗАИМНАЯ ИНФОРМАЦИЯ

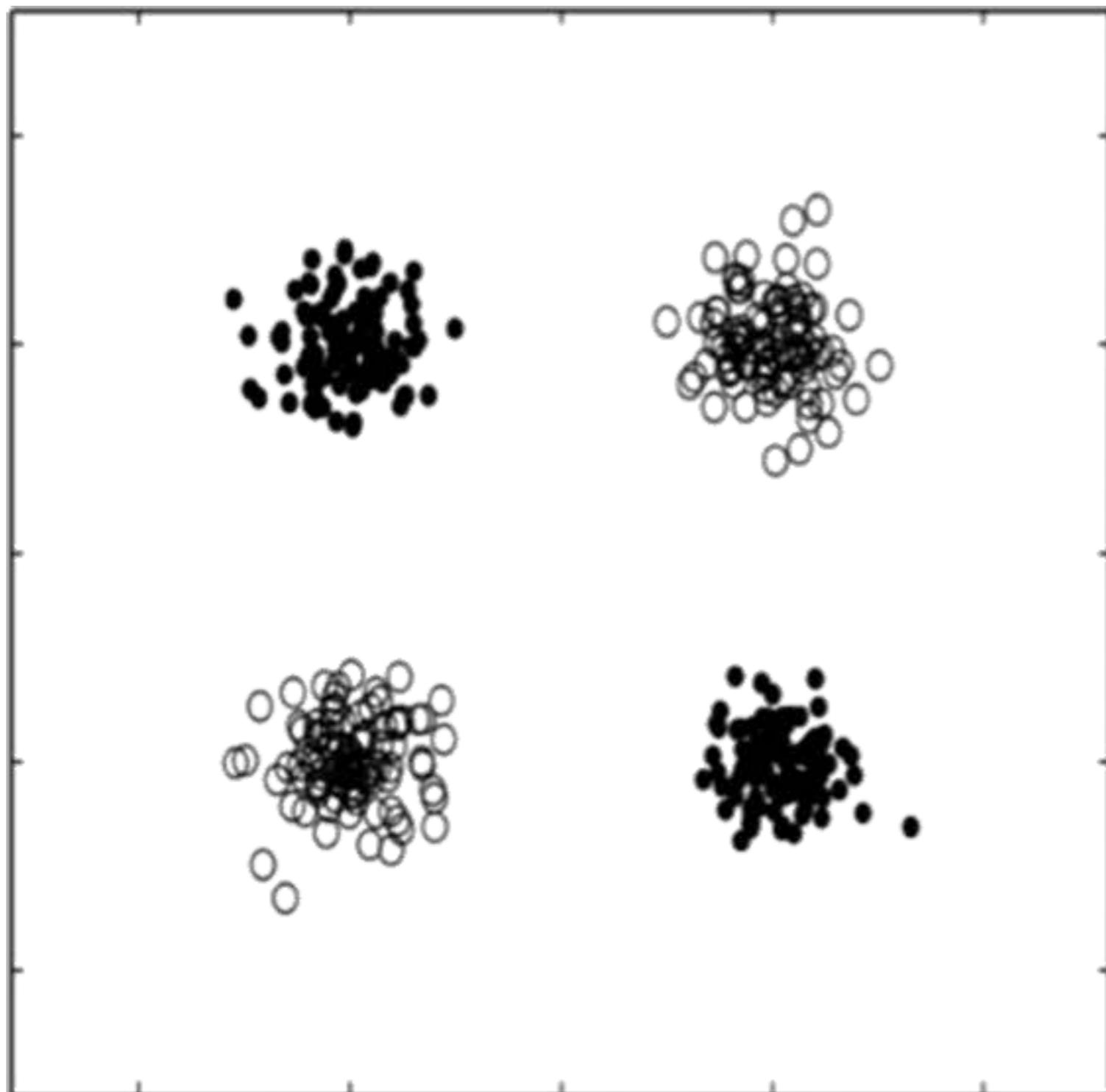
$$MI_j = \sum_{v=1}^n \sum_{k=1}^m P(x=v, y=k) \log \frac{P(x=v, y=k)}{P(x=v)P(y=k)}$$

- › Если признак и целевая переменная независимы, то $MI_j = 0$

ПРОБЛЕМЫ



ПРОБЛЕМЫ



РЕЗЮМЕ

- Одномерный отбор: корреляция, взаимная информация, **AUC**
- Не учитывает сложные зависимости

ЖАДНЫЕ МЕТОДЫ ОТБОРА ПРИЗНАКОВ

ОТБОР ПРИЗНАКОВ

➤ Надстройка над обучением:

- ▶ Перебираем комбинации признаков
- ▶ Для каждой обучаем модель
- ▶ Выбираем комбинацию, дающую лучшую модель

ПЕРЕБОРНЫЕ МЕТОДЫ

- Обучение модели — чёрный ящик
- Оптимизируем по подмножеству признаков
- Дискретная оптимизация
- Подойдут лишь методы, основанные на переборе вариантов



КАЧЕСТВО НАБОРА ПРИЗНАКОВ

- Выбираем подмножество признаков
- Обучаем модель только на них
- Оцениваем качество: отложенная выборка,
кросс-валидация

ПОЛНЫЙ ПЕРЕБОР

- › Пробуем все подмножества
- › Сначала размера 1, потом 2, 3, 4...

ПОЛНЫЙ ПЕРЕБОР

- › Пробуем все подмножества
- › Сначала размера 1, потом 2, 3, 4...
- › Порядка 2^d вариантов
- › Подходит для малого числа признаков

ЖАДНОЕ ДОБАВЛЕНИЕ

- Находим лучший набор из одного признака
 $J_1 = \{i_1\}$

ЖАДНОЕ ДОБАВЛЕНИЕ

- Находим лучший набор из одного признака

$$J_1 = \{i_1\}$$

- Находим признак, сильнее всего уменьшающий ошибку:

$$i_2 = \underset{i}{\operatorname{argmin}} Q(\{i_1, i\})$$

- Добавляем к набору: $J_2 = \{i_1, i_2\}$

ЖАДНОЕ ДОБАВЛЕНИЕ

- Находим лучший набор из одного признака

$$J_1 = \{i_1\}$$

- Находим признак, сильнее всего уменьшающий ошибку:

$$i_2 = \underset{i}{\operatorname{argmin}} Q(\{i_1, i\})$$

- Добавляем к набору: $J_2 = \{i_1, i_2\}$

- Продолжаем, пока ошибка уменьшается

ЖАДНОЕ ДОБАВЛЕНИЕ

- Быстро
- Но слишком жадно

ADD-DEL

- Чуть менее жадный перебор

ADD-DEL

- › Добавляем признаки, пока ошибка уменьшается
- › Удаляем по одному признаку, пока ошибка уменьшается
- › Продолжаем стадии добавления и удаления, пока удаётся уменьшать ошибку

ADD-DEL

- Всё ещё жадный алгоритм
- Но может исправлять ошибки, сделанные в процессе перебора

РЕЗЮМЕ

- Отбор признаков с целью оптимизации качества — самый прямой подход к задаче
- Но приводит к дискретной задаче оптимизации
- Можно решать перебором
- Жадные методы

ОТБОР НА ОСНОВЕ МОДЕЛЕЙ

ЛИНЕЙНЫЕ МОДЕЛИ

$$a(\mathbf{x}) = \sum_{j=1}^d \mathbf{w}_j x^j$$

- › Если признаки масштабированы, то веса можно использовать как показатели информативности
- › Для повышения числа нулевых весов — L_1 -регуляризация

РЕШАЮЩИЕ ДЕРЕВЬЯ

- Поиск лучшего разбиения:

$$Q(X_m, j, t) = \frac{|X_l|}{|X_m|} H(X_l) + \frac{|X_r|}{|X_m|} H(X_r) \rightarrow \min_{j, t}$$

- $H(X)$ — критерий информативности (MSE, Джини, энтропийный)

РЕШАЮЩИЕ ДЕРЕВЬЯ

- › Чем сильнее уменьшили $H(X)$, тем лучше признак
- › Уменьшение критерия:

$$H(X_m) - \frac{|X_l|}{|X_m|} H(X_l) - \frac{|X_r|}{|X_m|} H(X_r)$$

- › R_j : просуммируем уменьшения по всем вершинам, где разбиение делалось по признаку j

КОМПОЗИЦИИ

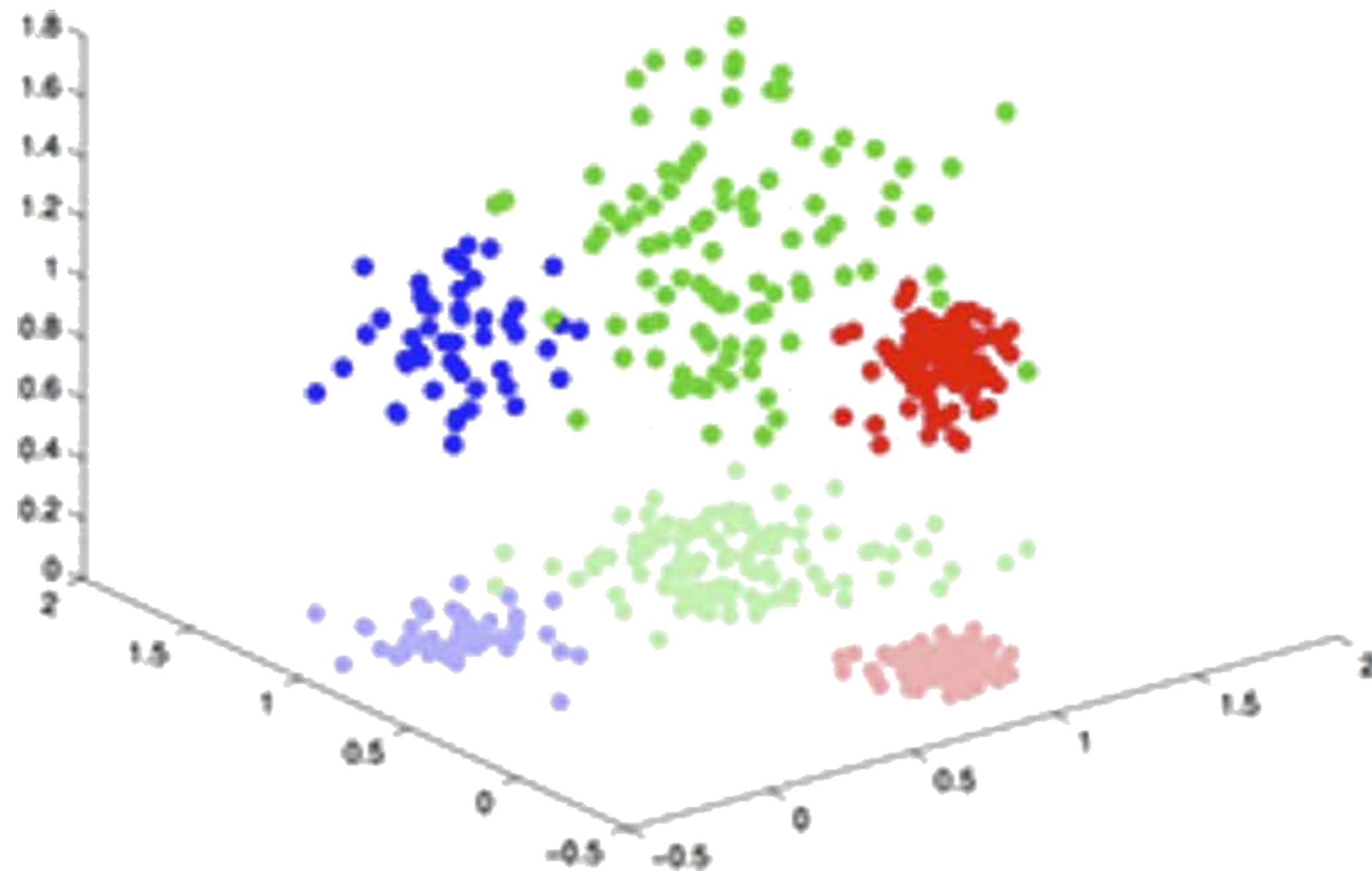
- Случайный лес и градиентный бустинг над деревьями
- Сумма R_j по всем деревьям
- Чем больше сумма, тем важнее признак

СЛУЧАЙНЫЙ ЛЕС

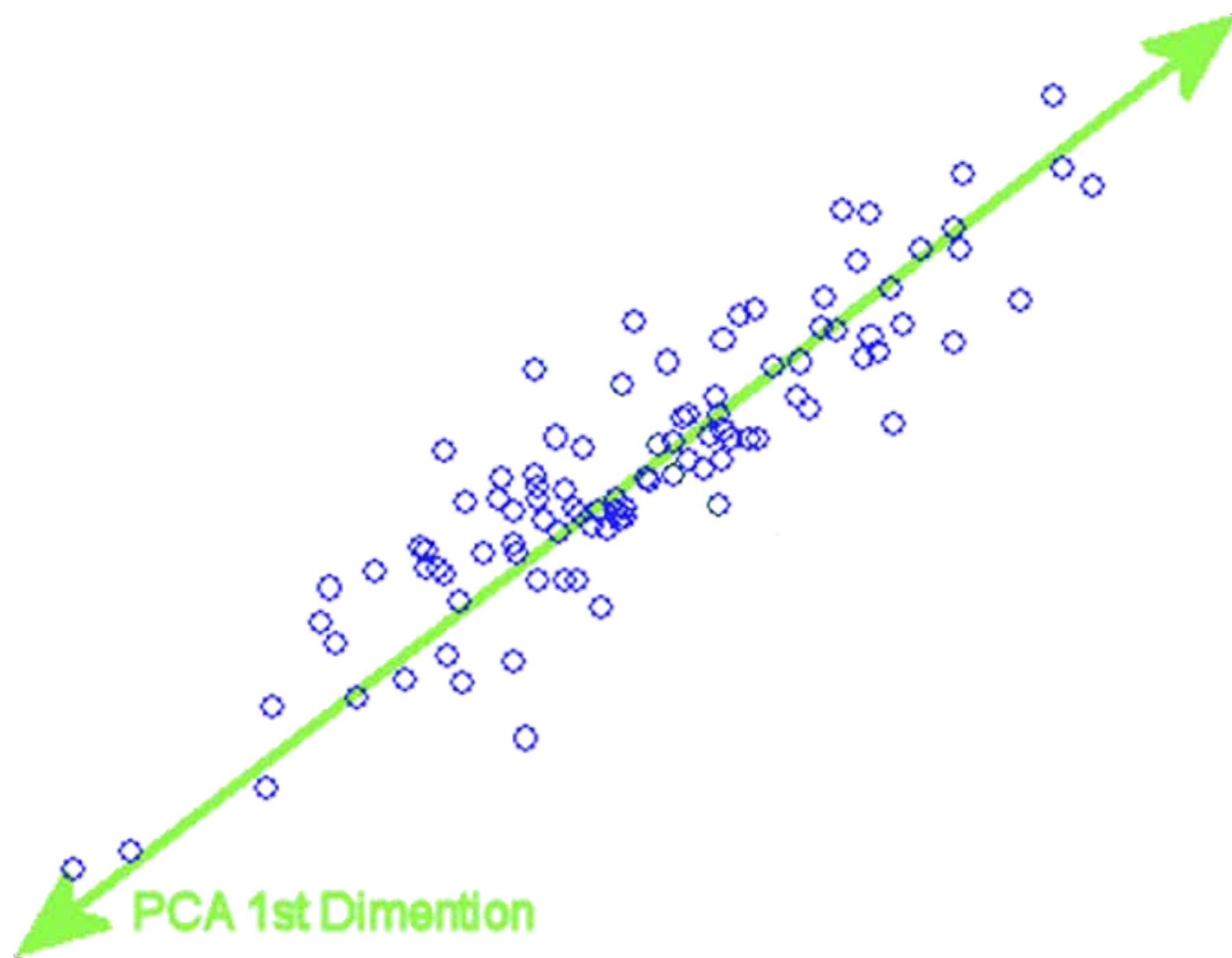
- Вычислим ошибку на Q_n out-of-bag-выборке для дерева b_n
- Переставим местами значения j -го признака
- Вычислим ошибку Q'_n на out-of-bag-выборке с переставленными значениями признака
- Информативность признака: $Q'_n - Q_n$
- Усредним $Q'_n - Q_n$ по всем деревьям

ПОНИЖЕНИЕ РАЗМЕРНОСТИ

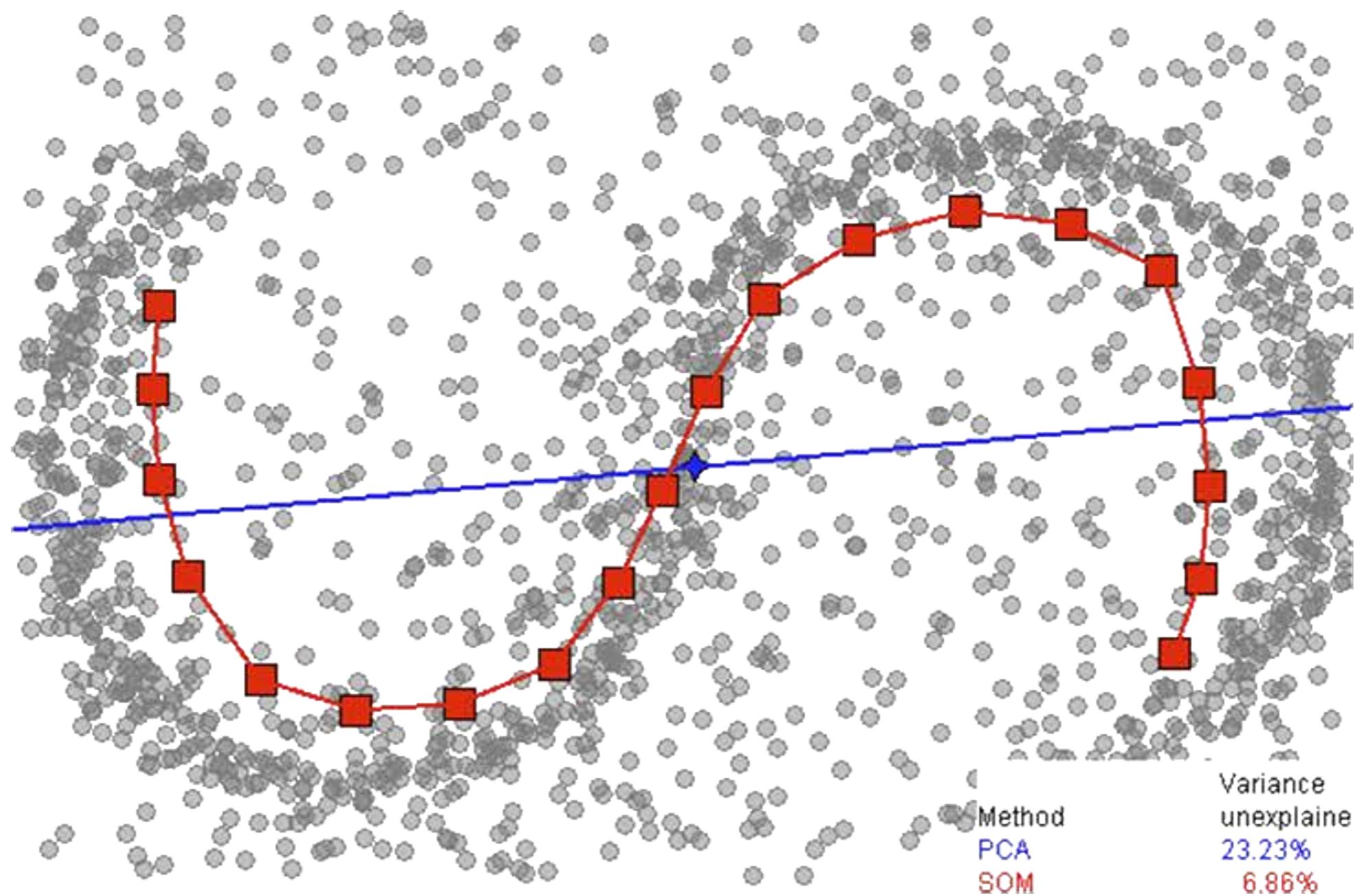
ПОНИЖЕНИЕ РАЗМЕРНОСТИ



ПОНИЖЕНИЕ РАЗМЕРНОСТИ



ПОНИЖЕНИЕ РАЗМЕРНОСТИ



ПОНИЖЕНИЕ РАЗМЕРНОСТИ

- › Порождение новых признаков
- › Их должно быть меньше
- › Они должны содержать как можно больше информации из исходных признаков

ПОНИЖЕНИЕ РАЗМЕРНОСТИ

- Линейные методы
- Каждый новый признак — линейная комбинация исходных

ПОНИЖЕНИЕ РАЗМЕРНОСТИ

- Исходные признаки: x_{ij} , D штук
- Новые признаки: z_{ij} , d штук
- Линейный подход: $z_{ij} = \sum_{k=1}^D w_{jk} x_{ik}$

ПОНИЖЕНИЕ РАЗМЕРНОСТИ

- › Исходные признаки: x_{ij} , D штук
- › Новые признаки: z_{ij} , d штук
- › Линейный подход:

$$z_{ij} = \sum_{k=1}^D w_{jk} x_{ik}$$

Новые признаки

Исходные признаки

Вклад исходного k -го
признака в новый j -й

МЕТОД СЛУЧАЙНЫХ ПРОЕКЦИЙ

› Линейный подход:

$$z_{ij} = \sum_{k=1}^D w_{jk} x_{ik}$$

› Выбираем веса случайно:

$$w_{ij} \sim \mathcal{N}(0, \frac{1}{d})$$

ЛЕММА ДЖОНСОНА-ЛИНДЕНШТРАУССА

➤ Идея:

- ▶ Если в выборке мало объектов и много признаков
 - ▶ То её можно спроектировать в пространство меньшей размерности
 - ▶ Так, что расстояния между объектами слабо изменятся
-
- Чтобы расстояния изменились не больше, чем на ϵ , надо взять $d > \frac{8 \ln \ell}{\epsilon^2}$

РЕЗЮМЕ

- › Понижение размерности
- › Простейший подход: метод случайных проекций

МЕТОД ГЛАВНЫХ КОМПОНЕНТ: ПОСТАНОВКА ЗАДАЧИ

ПОНИЖЕНИЕ РАЗМЕРНОСТИ

- › Исходные признаки: x_{ij} , D штук
- › Новые признаки: z_{ij} , d штук
- › Линейный подход:

$$z_{ij} = \sum_{k=1}^D w_{jk} x_{ik}$$

Новые признаки

Исходные признаки

Вклад исходного k -го
признака в новый j -й

МАТРИЧНАЯ ЗАПИСЬ

$$z_{ij} = \sum_{k=1}^D w_{jk} x_{ik} = \sum_{k=1}^D x_{ik} w_{kj}^T$$

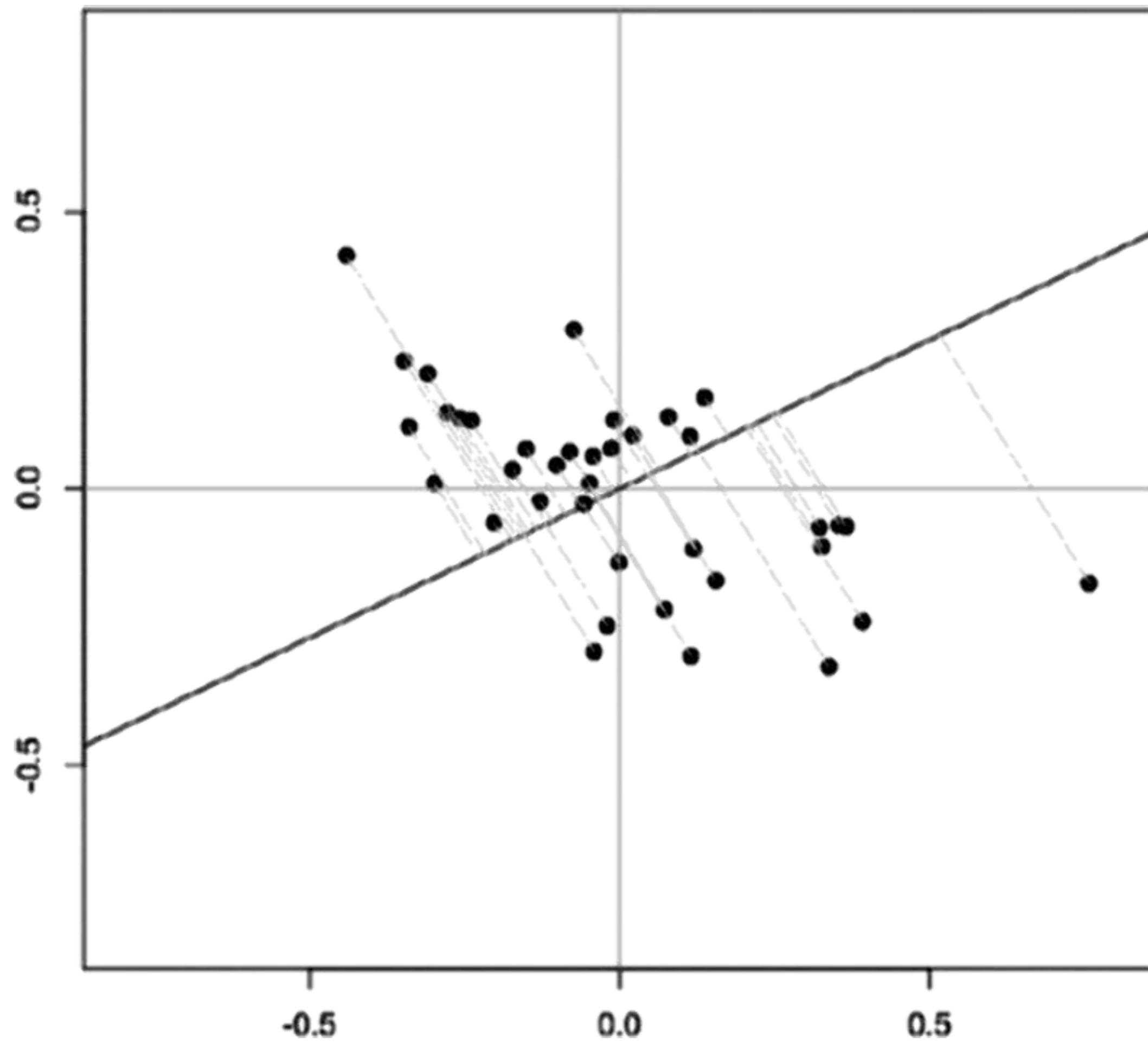


$$Z = XW^T$$

МАТРИЧНАЯ ЗАПИСЬ

- › Требование: $W^T W = I$
- › Тогда $X = ZW$
- › Задача: $\|X - ZW\|^2 \rightarrow \min_{Z, W}$

ПРОЕКЦИЯ НА ГИПЕРПЛОСКОСТЬ



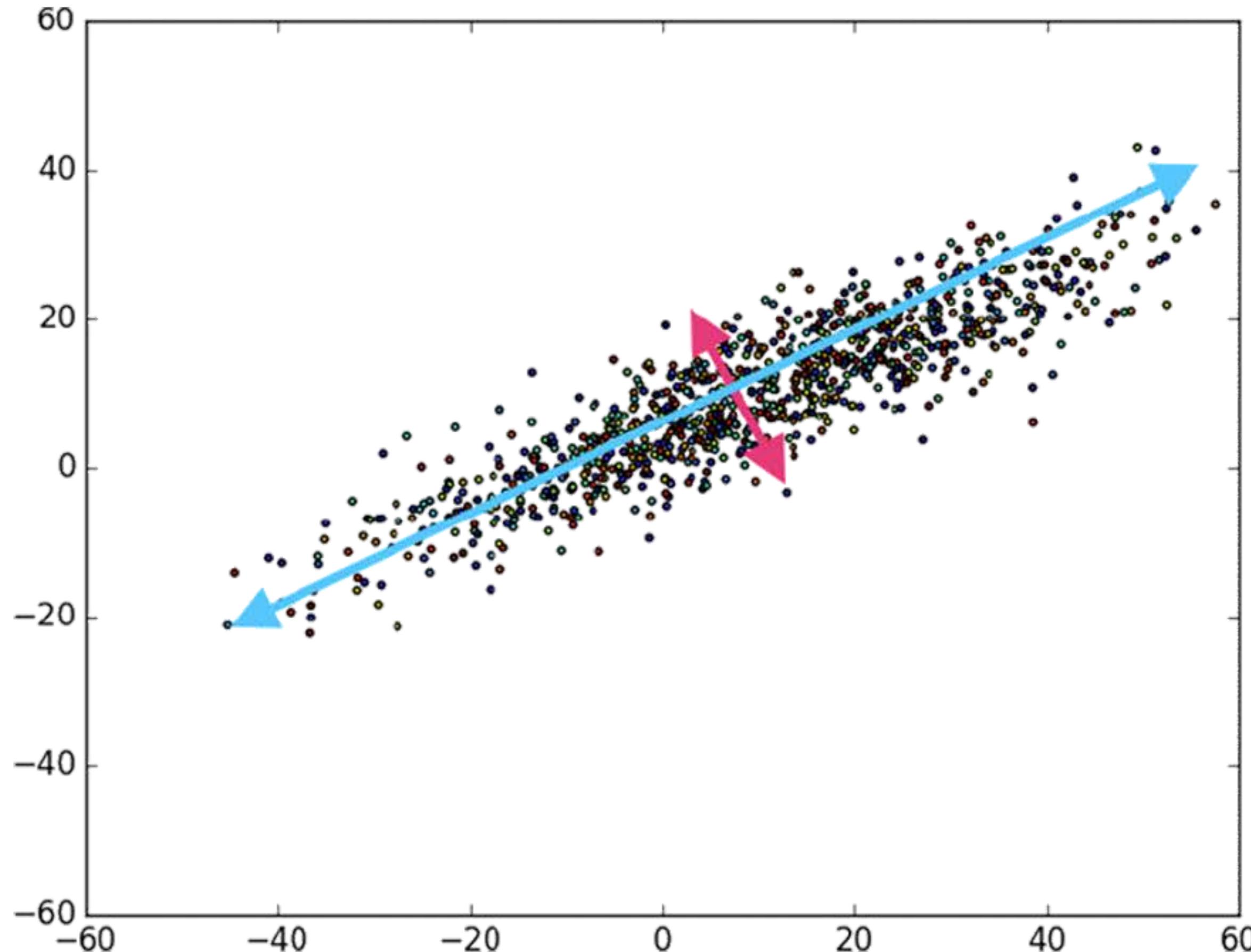
ПРОЕКЦИЯ НА ГИПЕРПЛОСКОСТЬ

$$\sum_{i=1}^{\ell} \|x_i - x_i W\|^2 \rightarrow \min_W$$



Проекция на плоскость,
столбцы W — направляющие
векторы

МАКСИМИЗАЦИЯ ДИСПЕРСИИ



МАКСИМИЗАЦИЯ ДИСПЕРСИИ

$$\sum_{j=1}^d \mathbf{w}_j^T \mathbf{X}^T \mathbf{X} \mathbf{w}_j \rightarrow \max_W$$

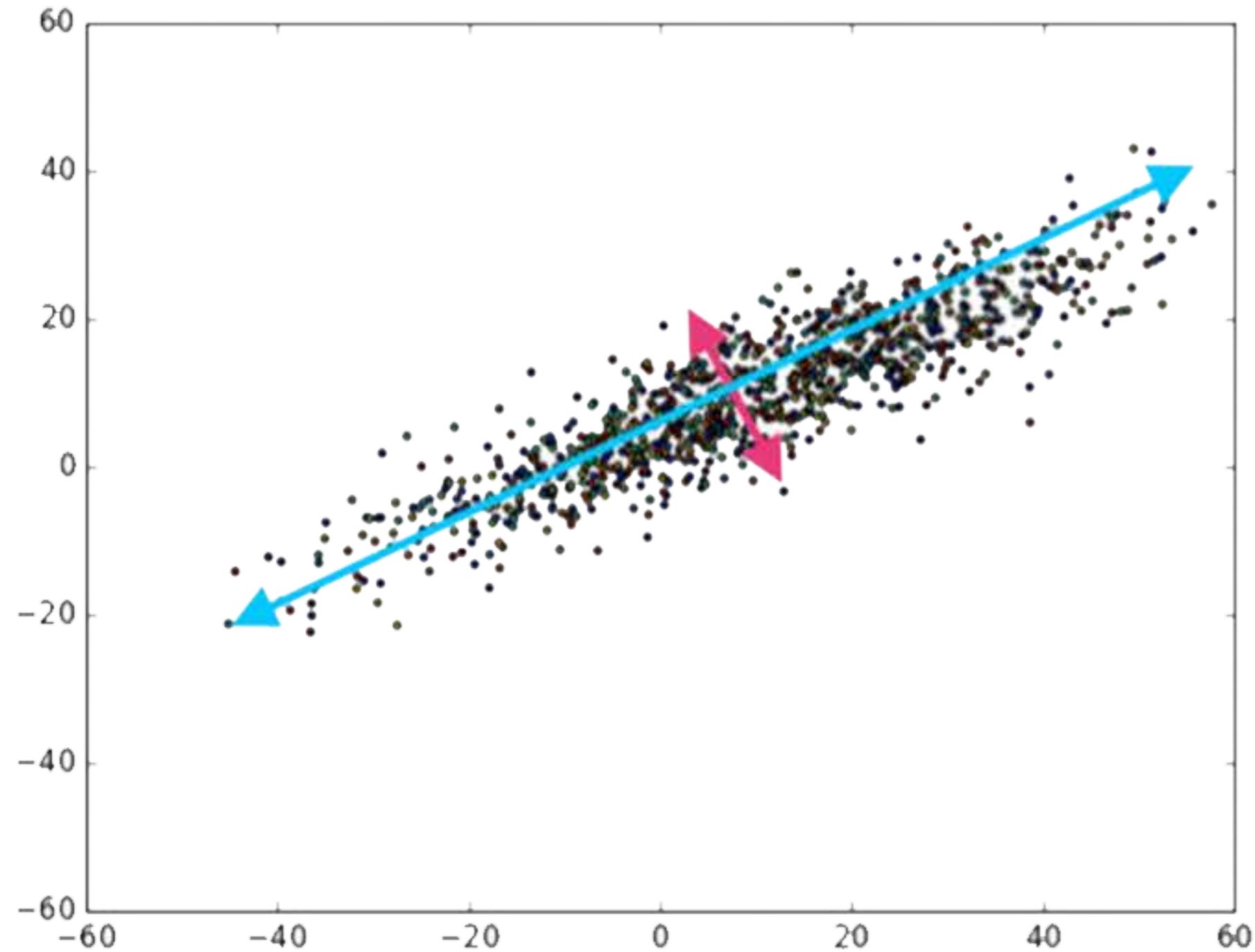
- › Чем больше, тем выше дисперсия выборки

РЕЗЮМЕ

- Матричное разложение
- Поиск проекционной гиперплоскости
- Максимизация дисперсии после проецирования
- Все постановки приводят к одному решению

МЕТОД ГЛАВНЫХ КОМПОНЕНТ: РЕШЕНИЕ

МАКСИМИЗАЦИЯ ДИСПЕРСИИ



МАКСИМИЗАЦИЯ ДИСПЕРСИИ

$$\begin{cases} \sum_{j=1}^d \mathbf{w}_j^T \mathbf{X}^T \mathbf{X} \mathbf{w}_j \rightarrow \max_{\mathbf{W}} \\ \mathbf{W}^T \mathbf{W} = \mathbf{I} \end{cases}$$

МАКСИМИЗАЦИЯ ДИСПЕРСИИ

$$\left\{ \sum_{j=1}^d \mathbf{w}_j^T \mathbf{X}^T \mathbf{X} \mathbf{w}_j \rightarrow \max_{\mathbf{W}} \right.$$

$$\left. \mathbf{W}^T \mathbf{W} = I \right.$$

Дисперсия
выборки

ЦЕНТРИРОВАНИЕ ВЫБОРКИ

- › $\sum_{j=1}^d \mathbf{w}_j^T \mathbf{X}^T \mathbf{X} \mathbf{w}_j$ — дисперсия только при условии, что выборка центрирована
- › Будем считать, что выборка центрирована — из каждого столбца вычли среднее

ПЕРВАЯ КОМПОНЕНТА

$$\begin{cases} \mathbf{w}_1^T X^T X \mathbf{w}_1 \rightarrow \max_{\mathbf{w}_1} \\ \mathbf{w}_1^T \mathbf{w}_1 = I \end{cases}$$

➤ \mathbf{w}_1 — вектор размера D

ПЕРВАЯ КОМПОНЕНТА

$$\begin{cases} \mathbf{w}_1^T X^T X \mathbf{w}_1 \rightarrow \max_{\mathbf{w}_1} \\ \mathbf{w}_1^T \mathbf{w}_1 = I \end{cases}$$

» Лагранжиан:

$$L(\mathbf{w}_1, \lambda) = \mathbf{w}_1^T X^T X \mathbf{w}_1 - \lambda(\mathbf{w}_1^T \mathbf{w}_1 - 1)$$

ПЕРВАЯ КОМПОНЕНТА

› Лагранжиан:

$$L(\mathbf{w}_1, \lambda) = \mathbf{w}_1^T \mathbf{X}^T \mathbf{X} \mathbf{w}_1 - \lambda (\mathbf{w}_1^T \mathbf{w}_1 - 1)$$

› Производная:

$$\frac{\partial L}{\partial \mathbf{w}_1} = 2 \mathbf{X}^T \mathbf{X} \mathbf{w}_1 - 2\lambda \mathbf{w}_1 = 0$$

ПЕРВАЯ КОМПОНЕНТА

› Лагранжиан:

$$L(\mathbf{w}_1, \lambda) = \mathbf{w}_1^T \mathbf{X}^T \mathbf{X} \mathbf{w}_1 - \lambda (\mathbf{w}_1^T \mathbf{w}_1 - 1)$$

› Производная:

$$\frac{\partial L}{\partial \mathbf{w}_1} = 2 \mathbf{X}^T \mathbf{X} \mathbf{w}_1 - 2\lambda \mathbf{w}_1 = 0$$

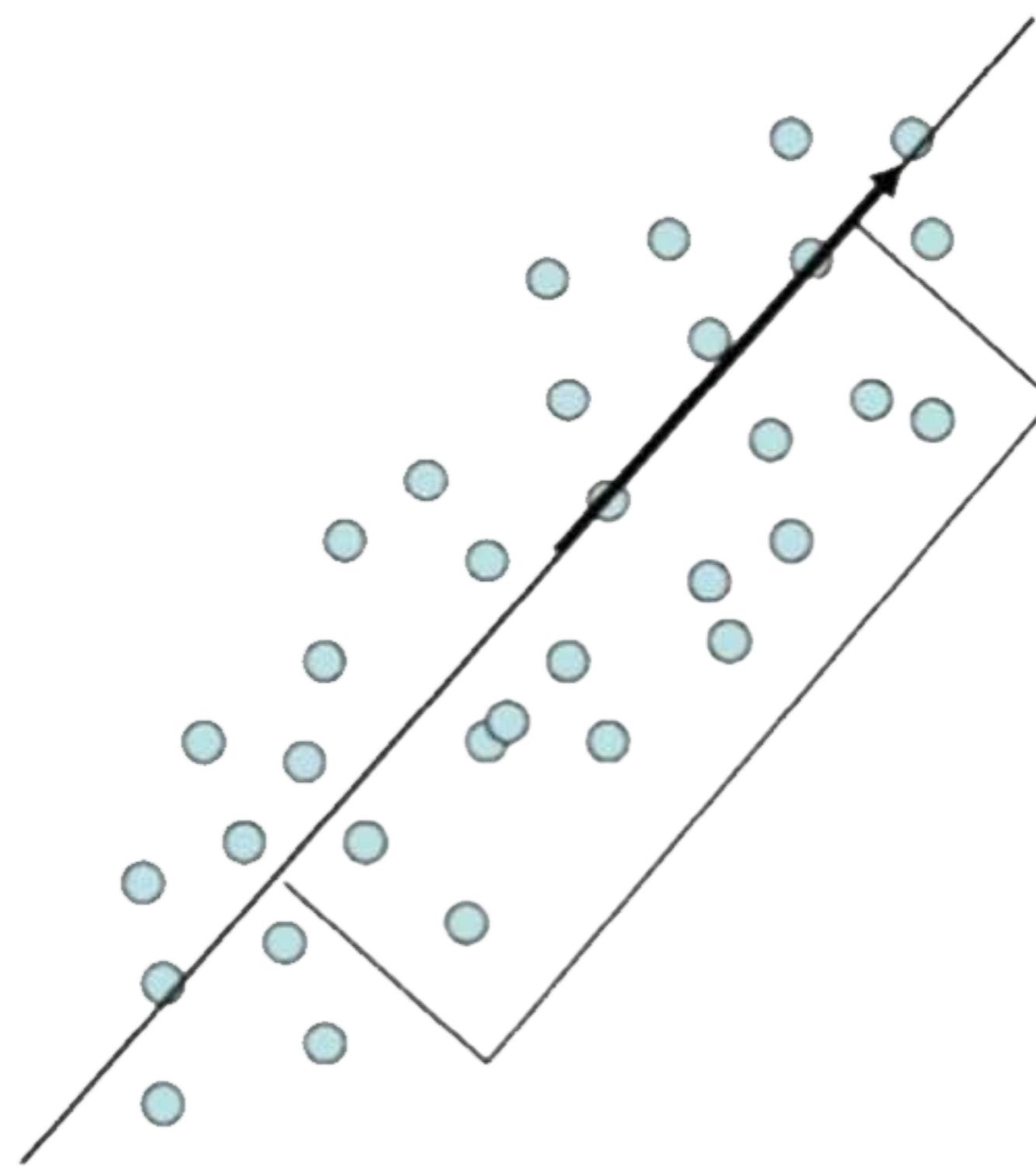
$$\mathbf{X}^T \mathbf{X} \mathbf{w}_1 = \lambda \mathbf{w}_1$$

ПЕРВАЯ КОМПОНЕНТА

$$\mathbf{X}^T \mathbf{X} \mathbf{w}_1 = \lambda \mathbf{w}_1$$

- › Дисперсия: $\mathbf{w}_1^T \mathbf{X}^T \mathbf{X} \mathbf{w}_1 = \lambda$
- › Первая компонента — собственный вектор матрицы $\mathbf{X}^T \mathbf{X}$, соответствующий наибольшему собственному значению
- › $\mathbf{X}^T \mathbf{X}$ — ковариационная матрица

ПЕРВАЯ КОМПОНЕНТА



ОБЩЕЕ РЕШЕНИЕ

- Оптимальные векторы $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_d$ — собственные векторы матрицы $\mathbf{X}^T \mathbf{X}$, соответствующие наибольшим собственным значениям $\lambda_1, \lambda_2, \dots, \lambda_d$
- $\frac{\sum_{i=1}^d \lambda_i}{\sum_{i=1}^D \lambda_i}$ — доля дисперсии, сохранённой при понижении размерности

СИНГУЛЯРНОЕ РАЗЛОЖЕНИЕ

- » $X = UDV^T$
- » Столбцы U — собственные векторы XX^T
- » Столбцы V — собственные векторы X^TX
- » Диагональ D — ненулевые собственные значения XX^T и X^TX (сингулярные числа)

МЕТОД ГЛАВНЫХ КОМПОНЕНТ

- › Найти сингулярное разложение матрицы X
- › Сформировать матрицу W из столбцов V , соответствующих наибольшим сингулярным числам
- › Преобразование признаков: $Z = XW$

РЕЗЮМЕ

- Оси для проекции в методе главных компонент — собственные векторы ковариационной матрицы
- Доля сохранённой дисперсии выражается через сингулярные числа