Московский физико-технический институт, МФТИ

Москва

Основные понятия и обозначения

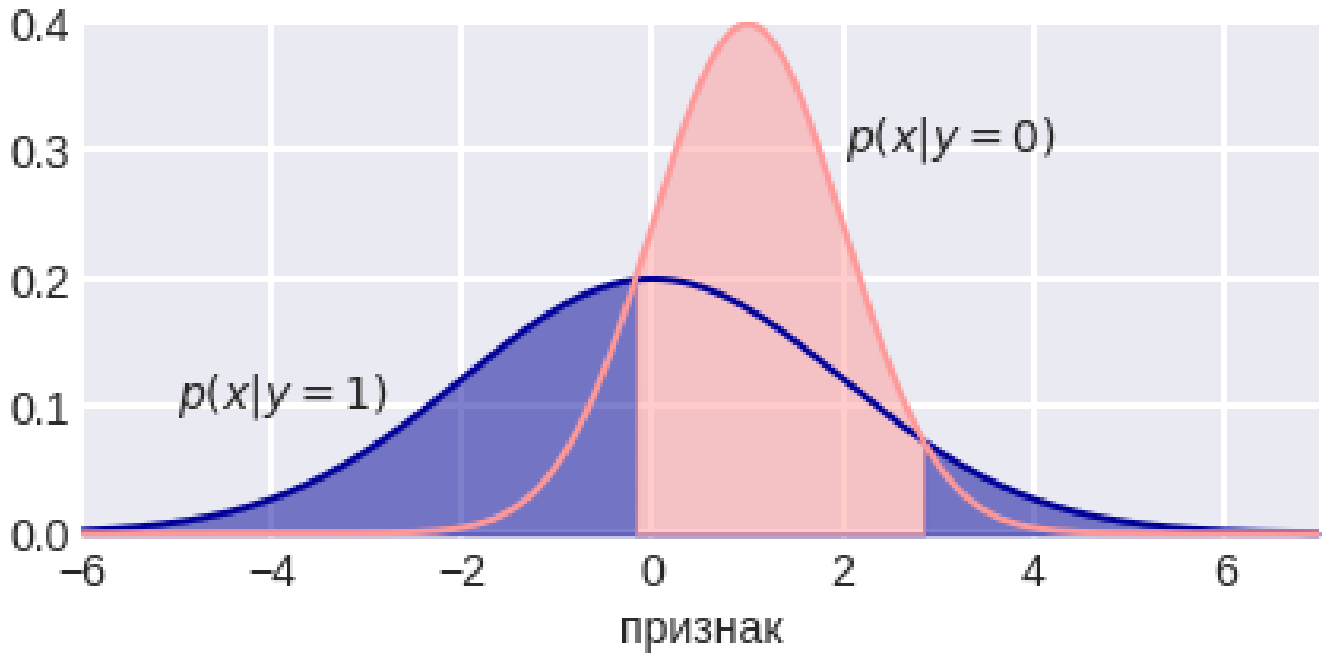
Задача классификаци X, Y, $X^l = (x_i, y_i)_{i=1}^l$ - выборка обучающая.
Необходимо построить неизвестное отображение: $a(x; X^l)$
$X \to Y$.
Формула Байеса:

$$p(x, y) = p(x)P(y|x) = P(y)p(x|y).$$

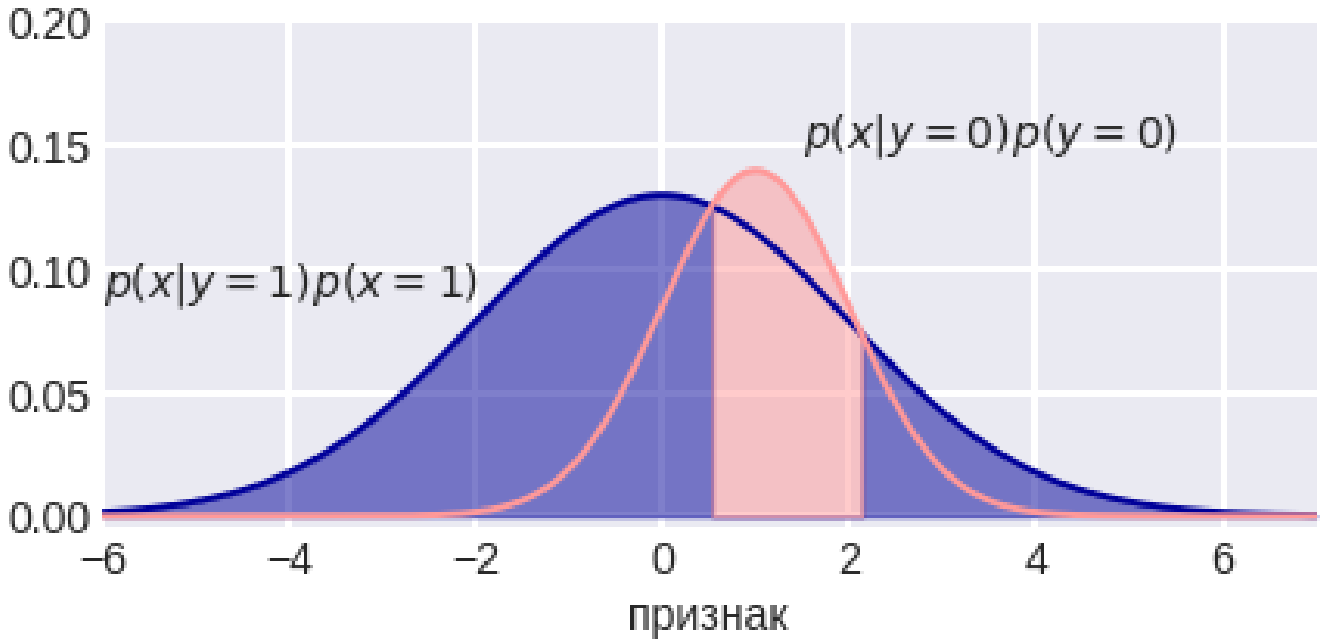
Принцип максимума апостериорной вероятности:

$$a(x) = arg \max_{y \in Y} P(y|x) = arg \max_{y \in Y} P(y)p(x|y).$$

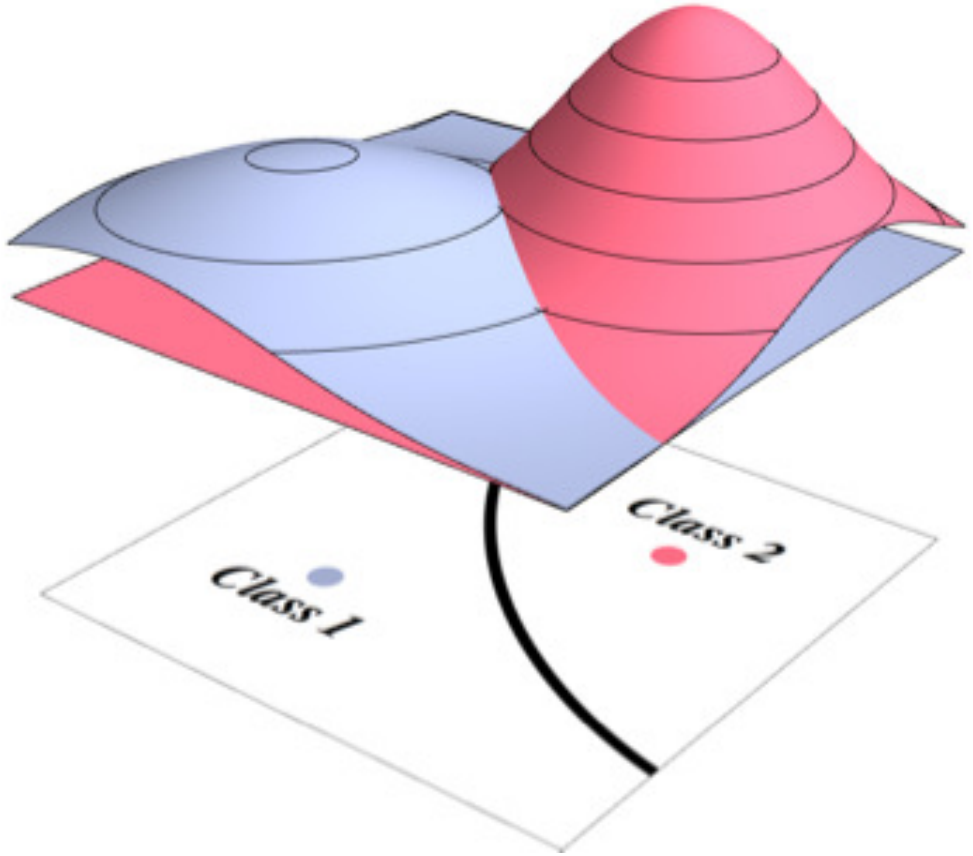
Распределение обьектов в каждом классе



Распределение обьектов в каждом классе, нормировка



Двумерное распределение обьектов



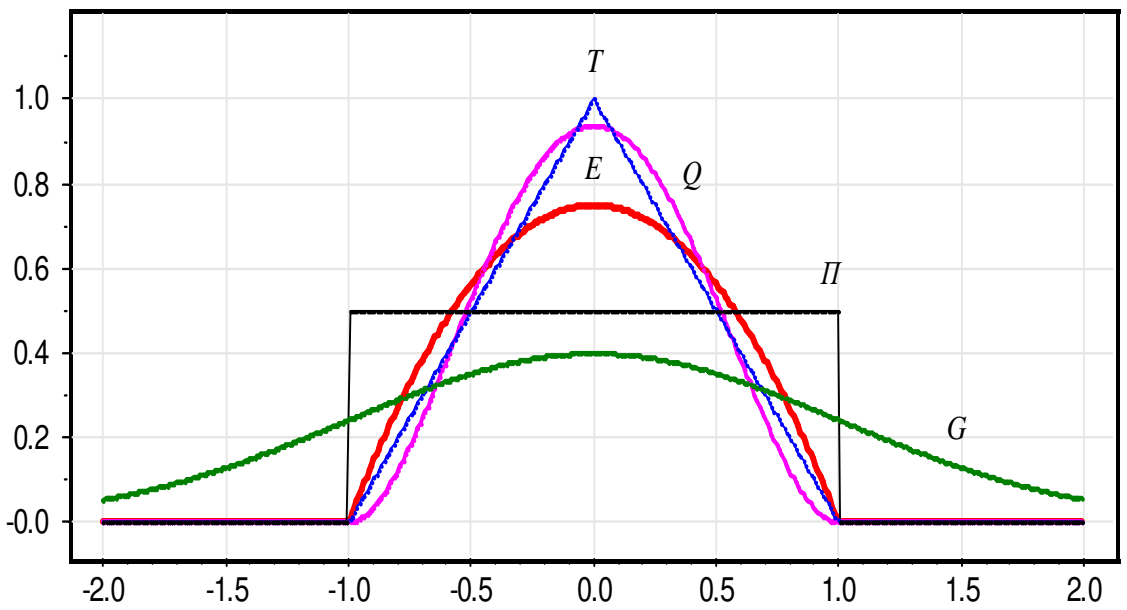
Способы оценки плотности распределений

- непараметрическая оценка плотности

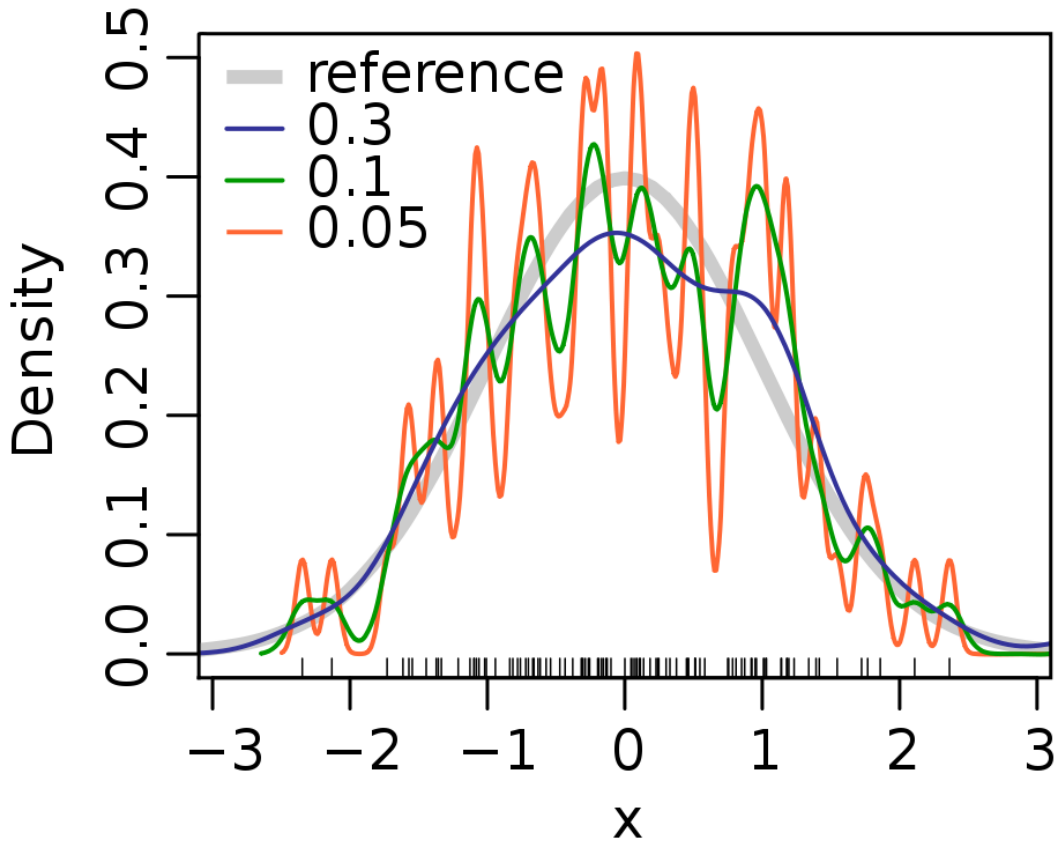
$$\hat{p}(x) = \sum_{i=1}^{n} \frac{1}{hV(h)} K\left(\frac{\rho(x, x_i)}{h}\right),$$

- параметрическая оценка плотности $\hat{p}(x) = \varphi(x, \theta)$;
- оценка смеси распределений $\hat{p}(x) = \sum w_j \varphi(x, \theta_j)$.

Функции в ядре



Пример непараметрического оценивания



Параметрическое оценивание

Допустим распределение обьектов оценивается гладкой функцией, зависящей от параметра θ:

$$p(x) = \varphi(x; \theta).$$

Принцип max правдоподобия

$$L(\theta; X^m) = \sum_{i=1}^{m} \ln \varphi(x_i; \theta) \to \max_{\theta}.$$

Распределение в классах

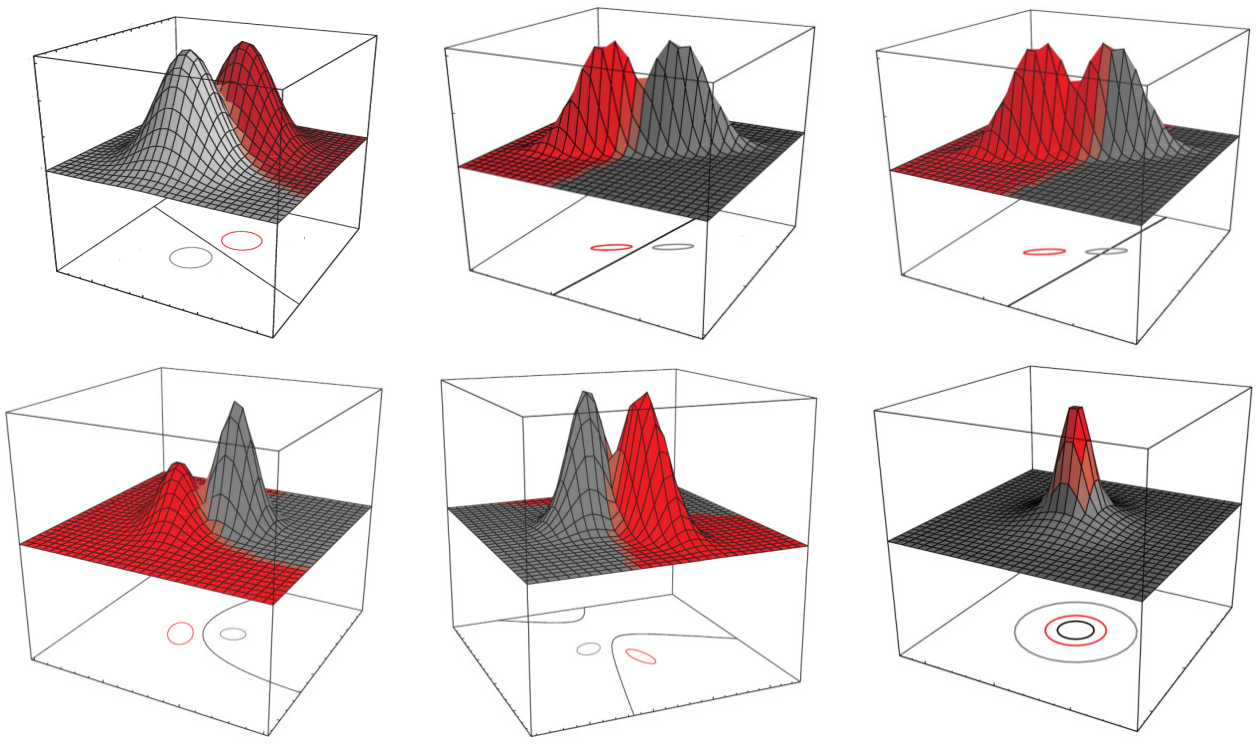
Рассмотрим многомерное гауссовское распределение:

$$p(x|y) = \mathcal{N}(x; \mu_y, \Sigma_y) = \frac{exp\{0.5(x-\mu_y)^T \Sigma_y^{-1}(x-\mu_y)\}}{\sqrt{(2\pi)^n det\Sigma_y}}.$$

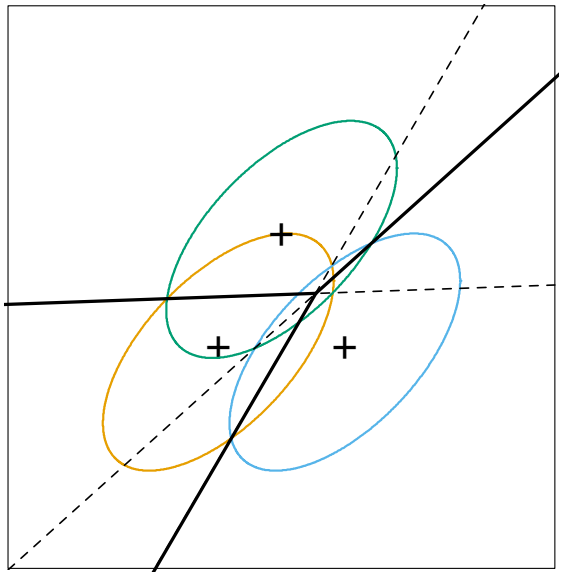
Разделяющая повернхность классов определятеся из соотношения

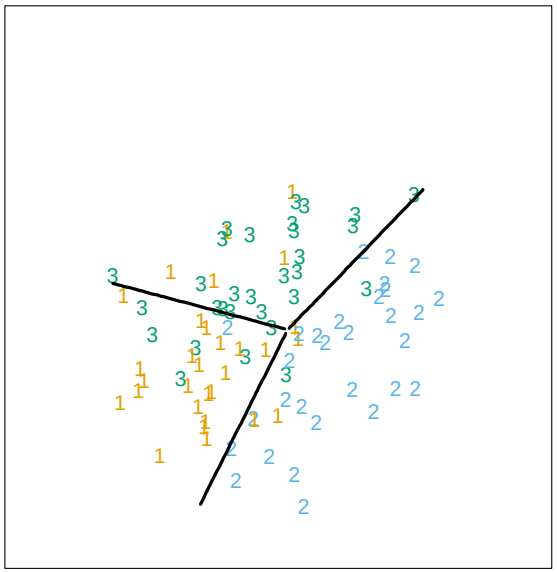
$$\lambda_y P(y)p(x|y) = \lambda_s P(s)p(x|s),\ y \neq s.$$

Примеры гауссовых двумерных распределений



Линейная граница нескольких классов





Линейный дискриминант Фишера

Пусть классы имеют одинаковую ковариацонную матрицу Σ, математическое ожидание обьектов отдельного класса:

$$\hat{\mu}_y = \frac{1}{l_y} \sum x_i,$$

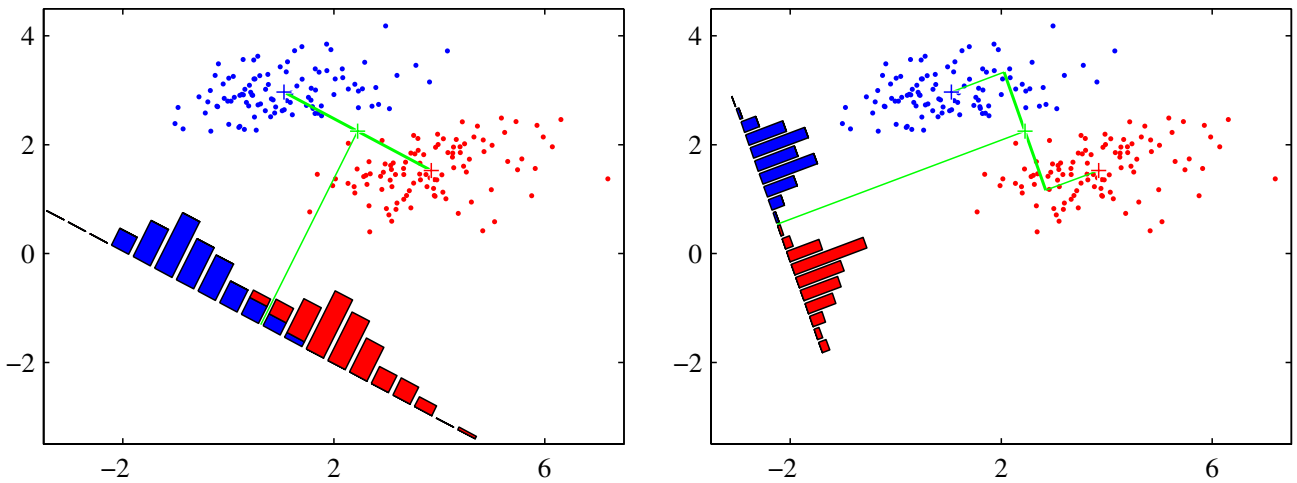
и среднее мат ожидание для двоих классов

$$\mu_{st} = 0.5(\mu_s + \mu_t),$$

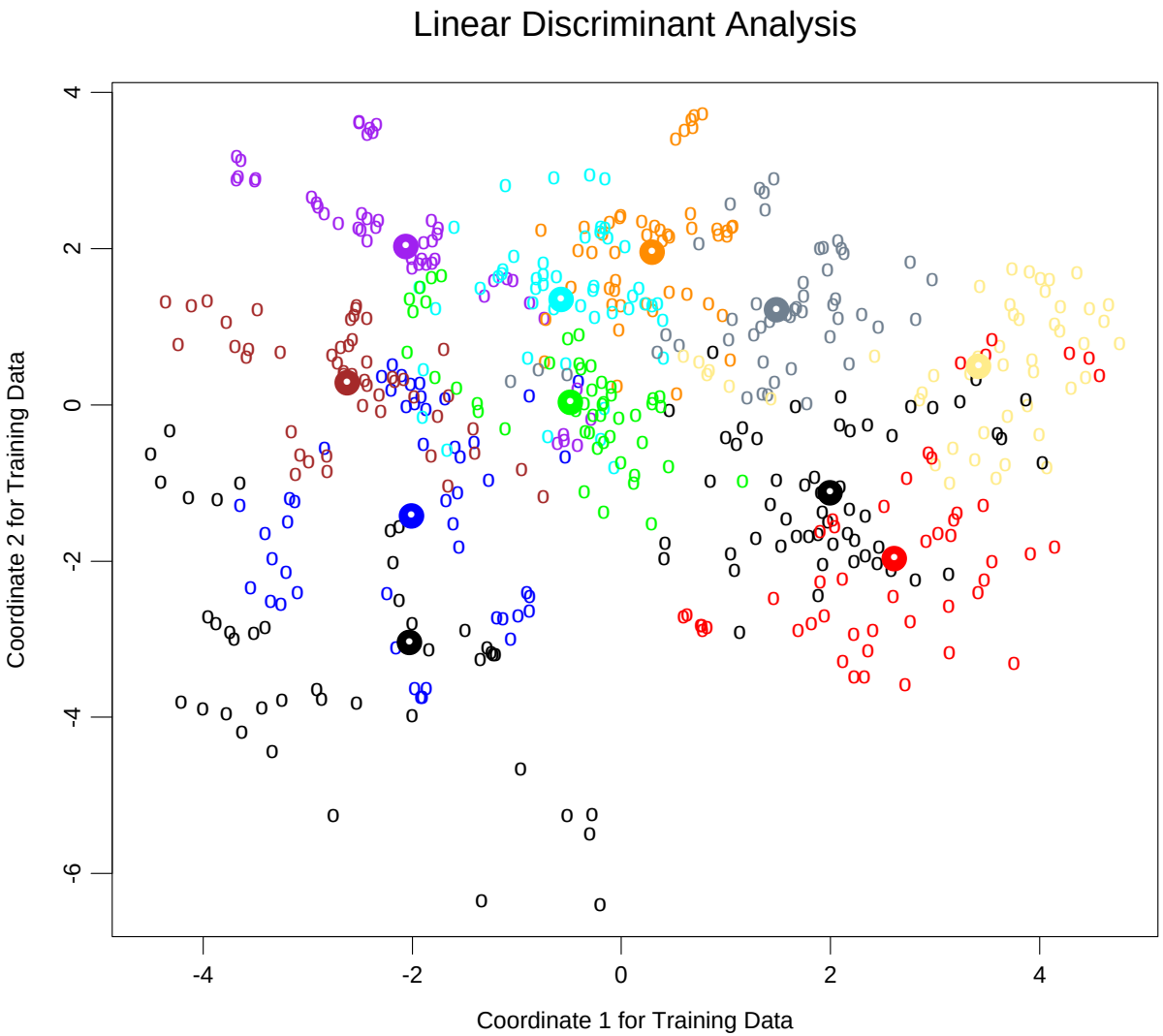
тогда граница классов определяется из уравнения

$$(x - \mu_{st})\Sigma^{-1}(\mu_s - \mu_t) = c_{st}.$$

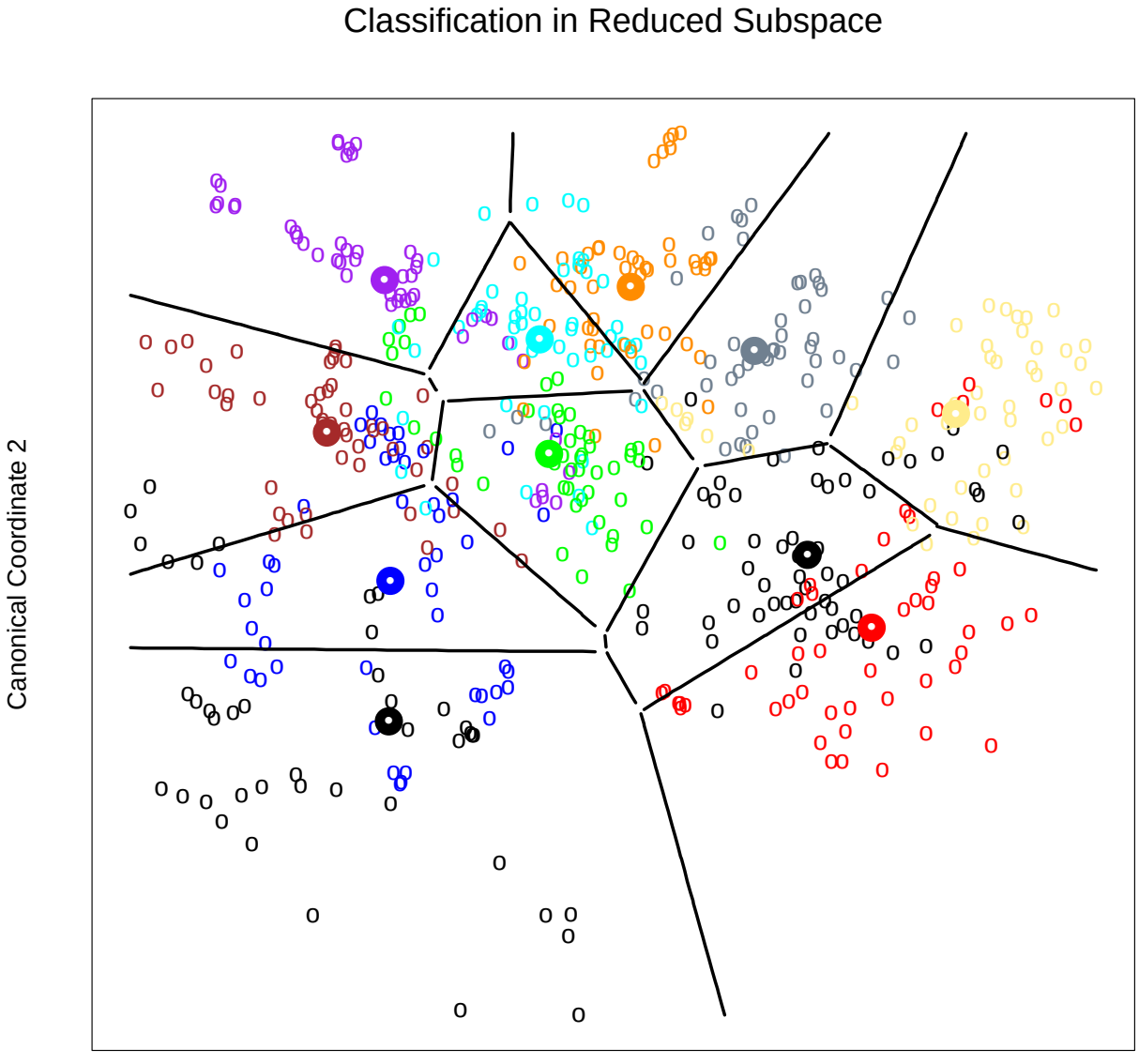
Линейный дискриминант Фишера



Пример множества классов



Пример границ разделения множества классов

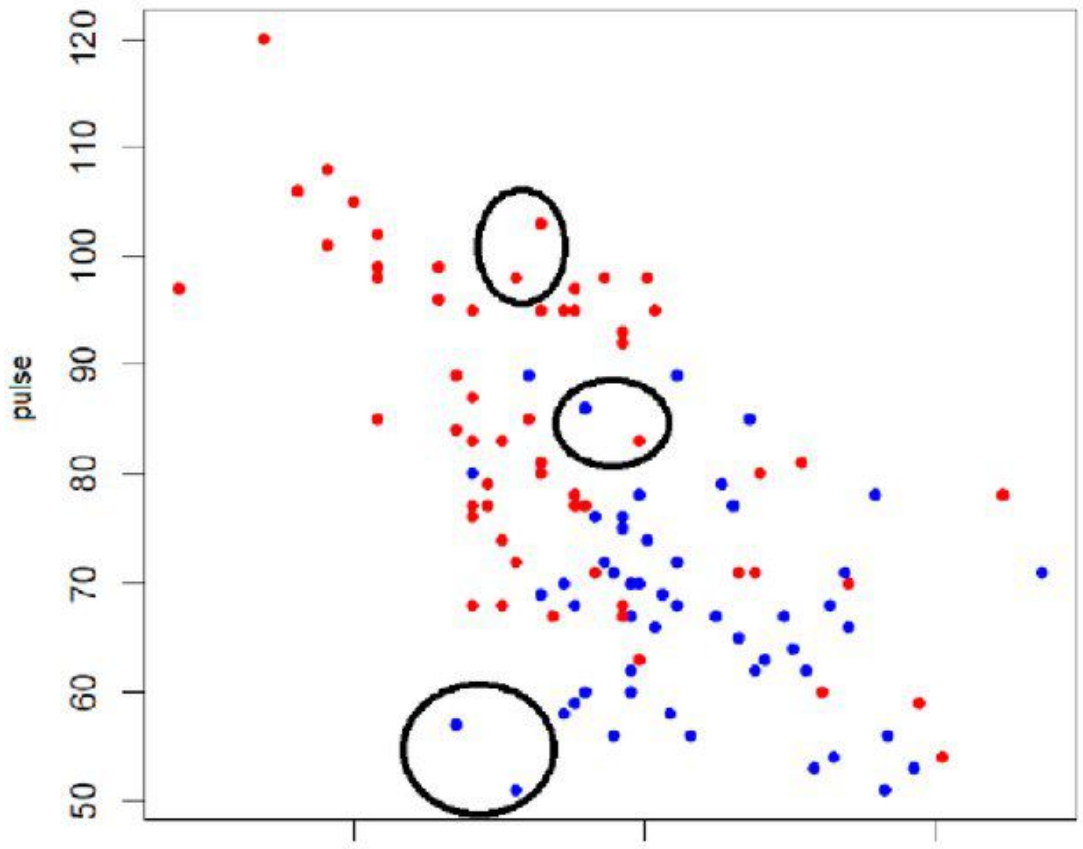


Метрическая классификация

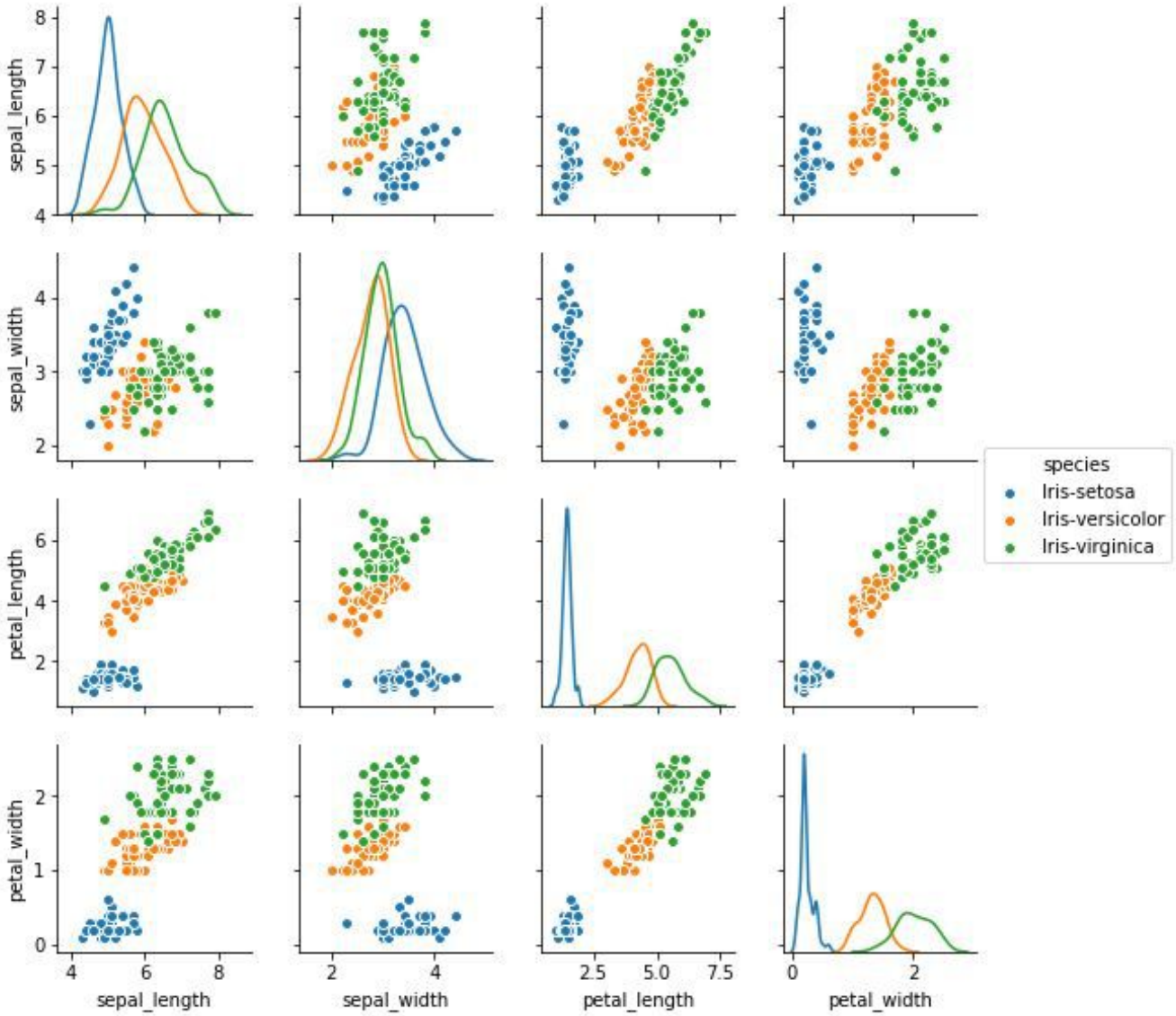
Задача классификаци X, Y, $X^l = (x_i, y_i)_{i=1}^l$ - выборка обучающая.

Возможно сравнить два отдельних элемента выборки в пространстве X по метрической функции $\rho(x_i, x_j)$.

Требуется найти алгоритм классификации $a(x, X^l) = Y$.



Иллюстрация, цветки ириса и др.



Различные виды норм

Рассмотрим функции описывающие различные нормы:

1. Евклидово $\rho(x, x_i) = \left(\sum_{j=1}^{n} w_k |x^j - x_i^j|^2\right)^{1/2}$;
2. L_p-метрика $\rho(x, x_i) = \left(\sum_{j=1}^{n} w_k |x^j - x_i^j|^p\right)^{p}$;
3. L_∞-метрика $\rho(x, x_i) = \max_{j=1..n} |x_i^j - x^j|$;
4. L_1-метрика $\rho(x, x_i) = \sum_{j=1}^{n} |x_i^j - x^j|$;

NB: $w_1, .., w_2$ - веса признаков, их также можно настраивать.

Обобщенный метрический классификатор

Метрический алгоритм классификации Пусть для заданной точки пространства обьектов X, его соседи из выборки X^l:

$$\rho(x, x^{(i)}) \leq \rho(x, x^{(i)}) \leq ... \leq \rho(x, x^{(l)}),$$

где $x^{(i)}$ - iй сосед обьекта x.

$$a(x; X^l) = \arg\max_{y \in Y} \sum_{i=1}^{l} [y^{(i)} = y] w(i, x),$$

так что можно определить функцию $\Gamma_y(x)$ - оценку близости обьекта x к классу y.

Метод ближайшего соседа

Алгоритм ближайшего соседа: $w(i,u) = [i = 1]$, $a(u; X^l) = y_u^{(1)}$.
Имеет ряд недостатков, такие как, неустойчивость, нельзя настроить под конкретные условия, необходимость хранить всю выборку.

Метод k средних kNN(k nearest neighbor)

Рассмотрим более широкую область вплость до k соседа $w(i,u) = [i \leq k]$. Настройка гиперпараметра k

$$LOO(k, X^l) = \sum_{i=1}^{l} [a(x_i; X^l \setminus x_i, k) \neq y_i] \to \min_k .$$

Одна из проблем - неоднозначность классификации:
$\Gamma_y(u) = \Gamma_s(u)$, $y \neq s$.

Метод взвешенных ближайших соседей

Возьмем коэффициенты с весами $w(i,u) = [i \leq k]w_i$.
Возможны следующие случаи выбора весов:

- $w_i = \frac{k+1-i}{k}$ - линейной убывающие веса;
- $w_i = q^i$ - экспонетоциально убывающие веса $0 < q < 1$;

Подбор параметра по LOO

Недостатки методов типа kNN

1. приходится хранить всю выборку целиком;
2. классифицируемый обьект сравнивается со всеми обьектами выборки $O(l)$, можно оптимизировать до $O(l \ln l)$ операций;
3. ограниченное число параметров и как следствие ограничения настройки по данным;

Метод парзеневского окна

Определим веса следующим образом $w(i,x) = K\left(\frac{\rho(x,x^{(i)})}{h}\right)$, h-ширина окна. Метод парзеневского окна фиксированной ширины

$$a(x; X^l, \boldsymbol{h}, K) = \arg\max_{y \in Y} \sum_{i=1}^{l} [y^{(i)} = y] K\left(\frac{\rho(x, x^{(i)})}{h}\right),$$

Метод парзеневского окна переменной ширины

$$a(x; X^l, \boldsymbol{k}, K) = \arg\max_{y \in Y} \sum_{i=1}^{l} [y^{(i)} = y] K\left(\frac{\rho(x, x^{(i)})}{\rho(x, x^{(k+1)})}\right),$$

Оптимизировать можно как параметры ширины окна h, k, так и вид самого ядра K.

Метод потенциальной функции

Воспользовавшись аналогией с зарядами возьмем следующие ядра:

$$w(i,x) = \gamma^{(i)} K\left(\frac{\rho(x,x^{(i)})}{h^{(i)}}\right),$$

Заметим, что обьекты можно не ранжировать

$$a(x;X^l) = \arg\max_{y\in Y} \sum_{i=1}^{l} [y^{(i)} = y]\gamma_i K\left(\frac{\rho(x,x_i)}{h_i}\right),$$

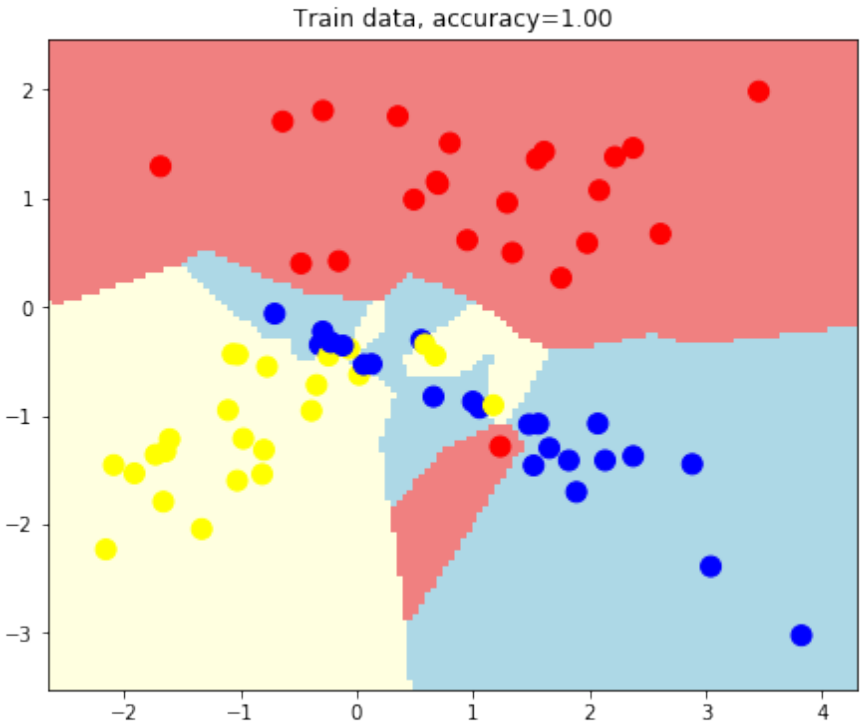
по физической аналогии: γ_i - величина заряда в точке x_i;
h_i - радиус действия потенциала с центом в точке x_i;
y_i - знак зараяда в случае двух классов $K(r) = \frac{1}{r}$.

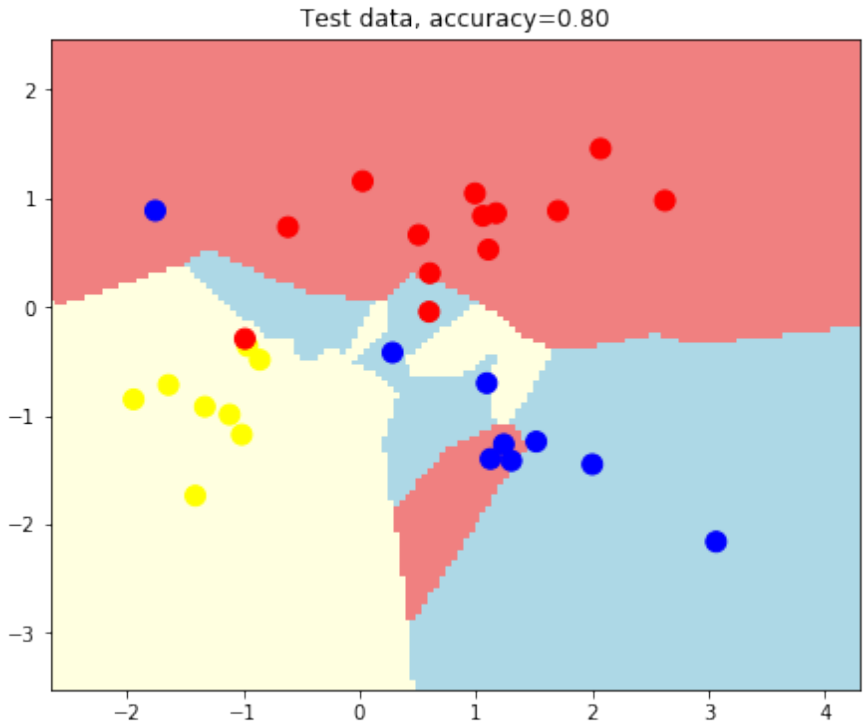
Достоинства и недостатки метода

Метрические классификаторы просты в реализации, позволяют с ходу решить множество задач.
При оптимизации возможно подбирать следующие параметры:

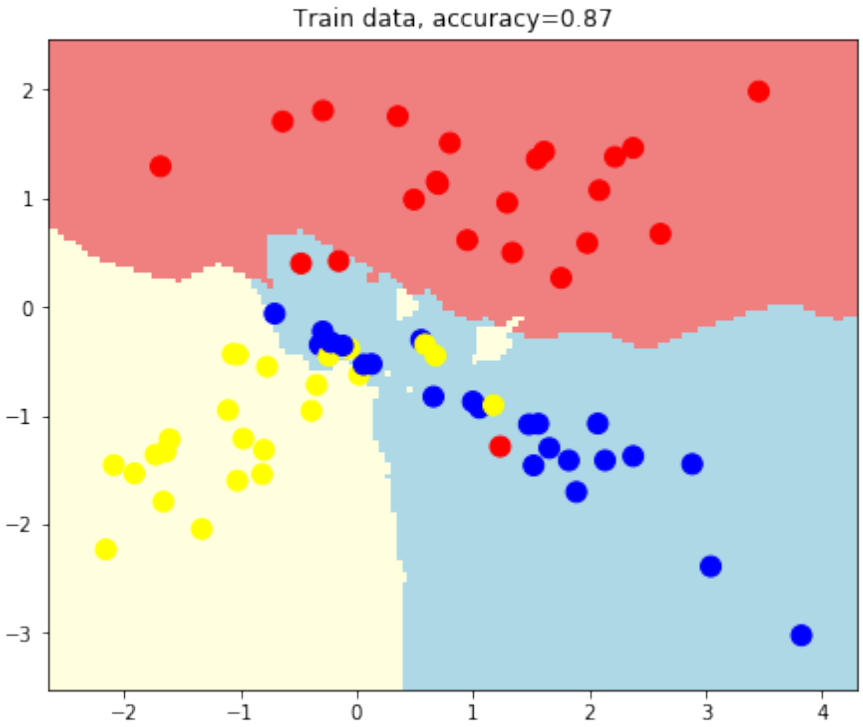
1. число соседей k и ширину окна h;
2. веса обьектов;
3. **набор эталонов**;
4. метрику (distance learning, similarity learning)
5. веса признаков самостоятельно;
6. функцию ядра $K(r)$.

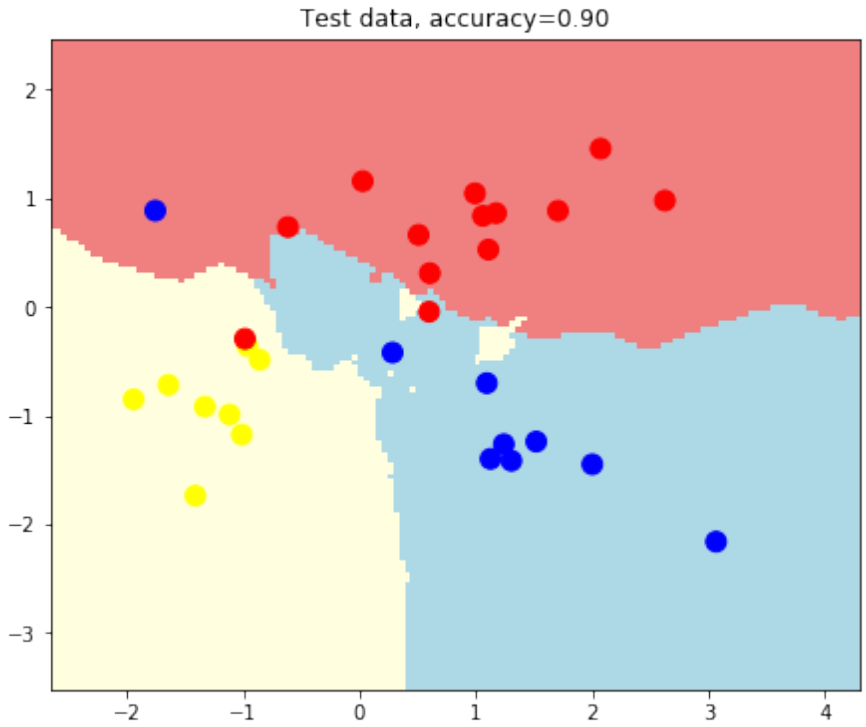
KNN 1


Train data, accuracy=1.00
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
4

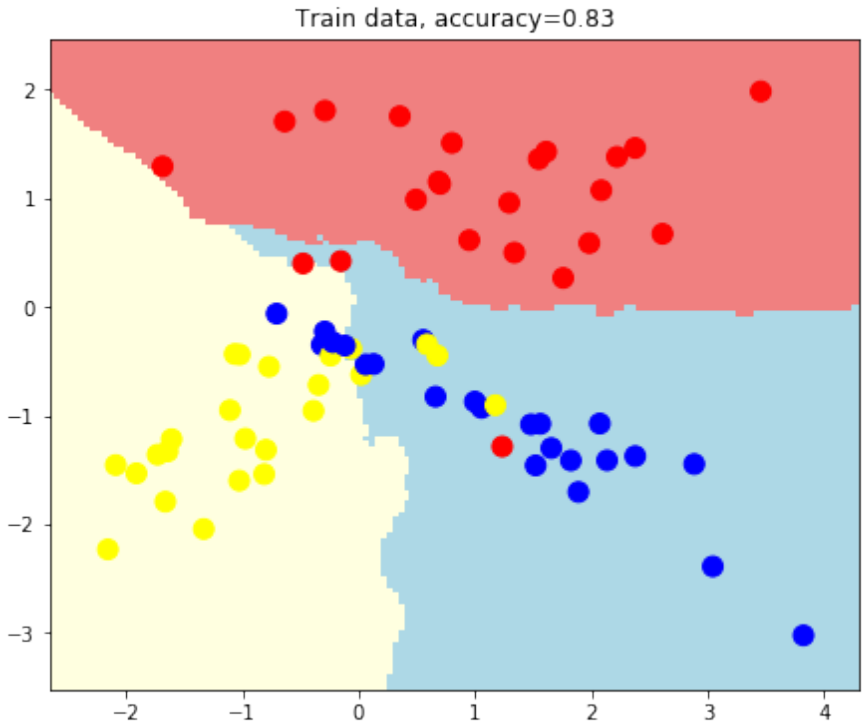

Test data, accuracy=0.80
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
4

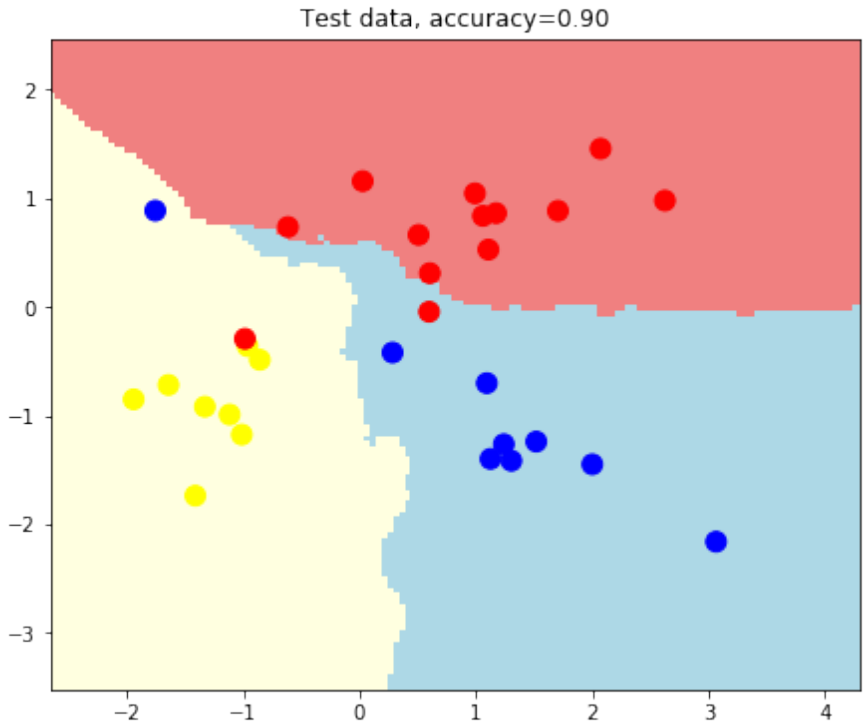
KNN 5


Train data, accuracy=0.87
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
4

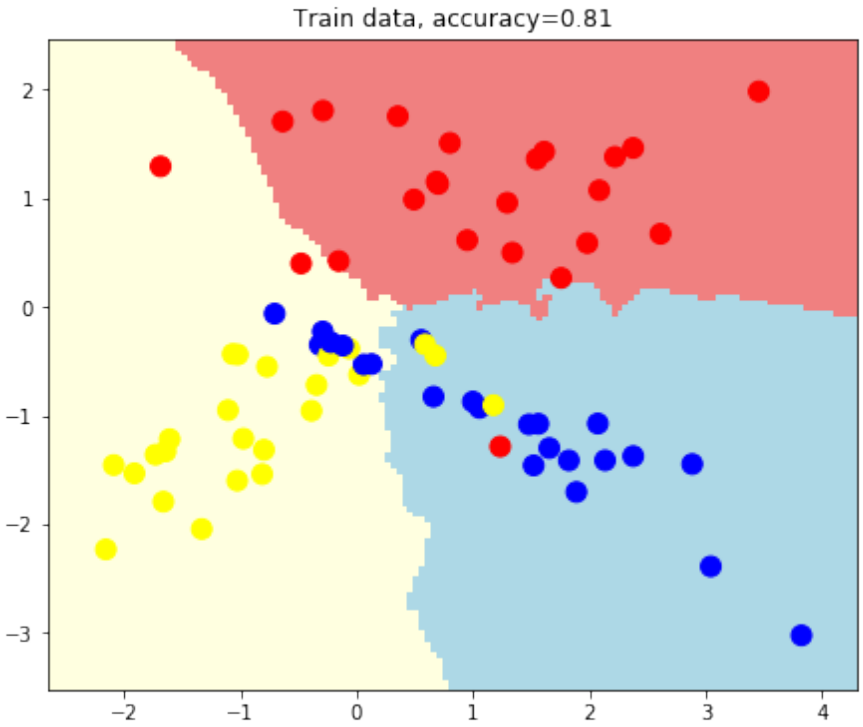

Test data, accuracy=0.90
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
4

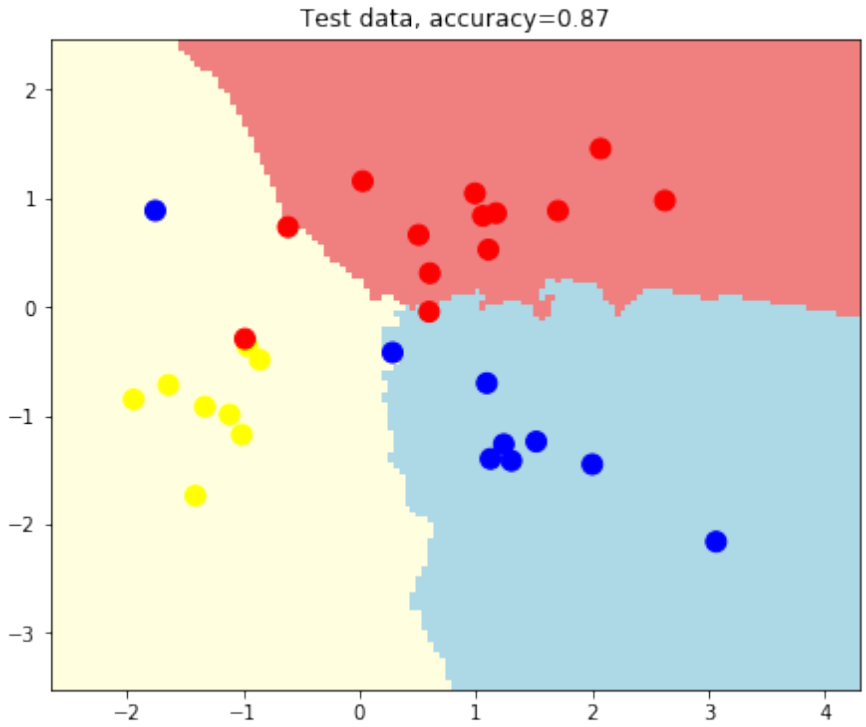
KNN 20


Train data, accuracy=0.83
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
4


Test data, accuracy=0.90
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
4

KNN 40


Train data, accuracy=0.81
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
4


Test data, accuracy=0.87
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
4

Отбор эталонных обьектов

Введем понятие "хороших"обьектов

1. эталон

различные обьекты негативно влияющие на результат:

1. неинформативные;
2. перифирийные;
3. шумовые выборсы;

Понятие отступа обьекта

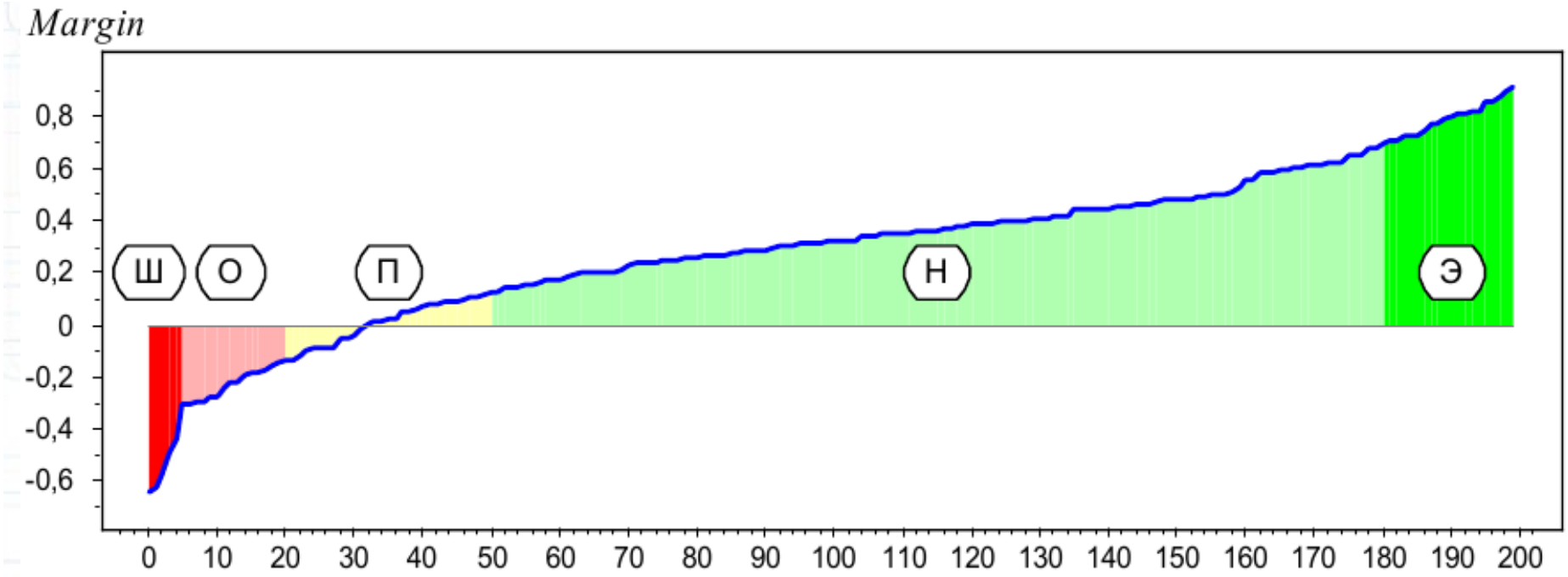
Определение

Отступом обьекта $x_i \in X^l$ *относительно классификации, имеющей вид* $a(u) = arg \max\limits_{y \in Y} \Gamma_y(u)$, *называется величина*

$$M(x_i) = \Gamma_{y_i}(x_i) - \max_{y \in Y \setminus y_i} \Gamma_y(x_i).$$

Отступ характеризует полезность обьекта . Отступ отрицателен тогда и только тогда, когда алгоритм допускает ошибку на данном обьекте.

Отступы обьектов, иллюстрация



Все типы обьектов относительно указанного обьекта можно разделить на группы:**эталонные, неинформативные, пограничные, ошибочные, шумовые.**

Алгоритм STOLP для отбора эталонных обьектов

Вход: X^l - обучающая выборка,δ - порог фильтрации выбросов, l_0 - допустимая доля ошибок.
Выход: Множество опорных обьектов $\Omega \subseteq X^l$.
Модель классификатора:

$$a(x;\Omega) = \arg\max_{y \in Y} \sum_{x_i \in \Omega}^{l} [y_u^{(i)} = y] w(i,x),$$

$y_u^{(i)}$ - ответ на i-м соседе обьекта u.

Алгоритм STOLP

1. Для всех $x_i \in X^l$ проверить, является ли он выбросом $(M(x_i, X^l) < \delta)$
2. $\Omega = \{\arg\max_{x_i \in X^l_y} M(x_i, X^l) | y \in Y\}$,

   пока $\Omega \neq X^l$
3. Выделить множество обьектов с ошибкой $a(u, \Omega)$:
   $E = \{x_i \in X^l \setminus \Omega : M(x_i, \Omega) < 0\}$
   если $|E| < l_0$, то выход
4. Присоединить к Ω обьект с наименьшим отступом:
   $x_i = \arg\min_{x \in E} M(x, \Omega)$, $\Omega = \Omega \cup x_i$.

Анализ алгоритма

Преимущества методов

1. уменьшаем число хранимых обьектов
2. время классификации
3. обьекты распределяются по величине отступов (см картинку, характер прямой - качество классификации)

Недостатки метода

1. дополнительный параметр δ
2. эффективность $O(|\Omega|^2 l)$

Выбор метрики

Метрика Минковского с весами:

$$\rho(x, x_i) = \left(\sum_{j=1}^{n} w_k |x^j - x_i^j|^p \right)^p$$

- нормировка признака;
- упорядочевание признаков по важности;
- отрбасывание признаков;

Жадное добавление признаков

1. Решим задачу по каждому из признаков $\rho_j(u, x_i) = |u^i - x_i^j|$. Выберем налиучший параметр $LOO(j) \to min$.
2. Расширим нашу метрику добавлением еще одного признака

$$\rho'(u, x_i) = \rho(u, x_i) + w_j \rho_j(u, x_i),$$

   одновременно найдем признак j и $w_j \geq 0$ наиболее оптимальные $LOO(j, w_j) \to min$ (два цикла).
3. жадный алгоритм может нахватать лишних признаков

$$\rho'(u, x_i) = \rho(u, x_i) - w_k \rho_k(u, x_i) + w_j \rho_j(u, x_i),$$

   (хорошо работает для правильных задач).
4. Будем добавлять признаки, пока LOO увеличиваться.

Полный скользящий контроль

Complete cross validation (CCV)

$$CCV(X^l) = \frac{1}{C_L^l} \sum_{X^l \sqcup X^k} \frac{1}{k} \sum_{x_i \in X^k} [a(x_i, X^l) \neq y_i],$$

где $X^l \sqcup X^k$ - все C_L^l разбиений выборки X^L.

Выписываем точную комбинарторную формулу для полного скользящего контроля.

При $k = 1$ $CCV(X^L) = LOO(X^L)$.

CCV характеризует среднюю часоту ошибки, не учитывая пространственое распределение обьектов.

Профиль компактности

Definition

Профиль компактности выборки X^L-это функция доли обьектов x_i, у которых m-сосед x_i^m лежит в другом классе

$$K(m, X^L) = \frac{1}{L}\sum_{i=1}^{L}[y_i \neq y_i^{(m)}],\ m = 1, .., L-1.$$

Можно получить формулу скользящего контроля в явном виде

$$CCV(X^L) = \sum_{m=1}^{k} K(m, X^L)\frac{C_{L-1-m}^{l-1}}{C_{L-1}^{l}}.$$

$$CCV(X^L) = \sum_{m=1}^{k} K(m, X^L) \frac{C_{L-1-m}^{l-1}}{C_{L-1}^{l}}.$$

- связь свойств выборки с качеством классификации, формализация гипотезы компактности;
- слабая зависимость от длины контроля;
- важен начальный участок профиля, в силу ассимптотики $\frac{C_{L-1-m}^{l-1}}{C_{L-1}^{l}} \to 0$ по m;
- применим для обора эталонов;

Заключение

- Метрическая классификация проста в реализации, универсальна;
- Различные варианты для обучения:
— число соседей,
— веса обьектов,
— набор эталонов,
— метрику,
— веса признаков;
- Распределение отступов дает возможность разбить обьекты на классы (эталонные, малоинформативные, пограничные, ошибочные и шумовые); - Профиль компактности позволяет адаптировать выборку;