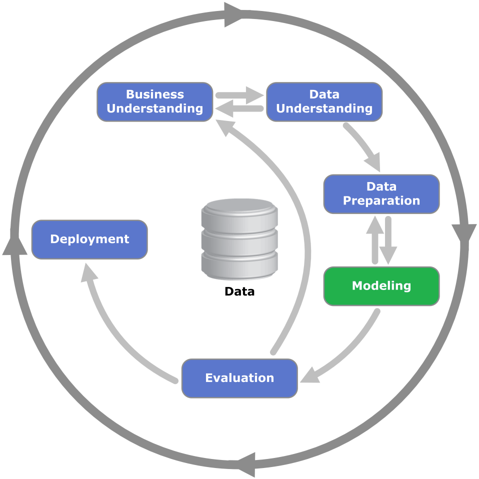
# Введение в Data Science
# Занятие 2. Линейные модели

Николай Анохин Михаил Фирулик

10 марта 2014 г.

ТЕХНОСФЕРА@mail.ru

# Где мы находимся (глобально)



# Где мы находимся (локально)

M Выдвигаем гипотезу насчет **модели** - семейства параметрических функций вида

$$Y = \{y(x, \theta) : X \times \Theta \to T\},$$

которая могла бы решить нашу задачу (model selection)

L Выбираем наилучшие параметры модели $\theta^*$, используя **алгоритм обучения**

$$A(\mathbf{X}, \mathbf{T}) : (X, T)^N \to Y$$

(learning/inference)

D Используя полученную модель $y^*(x) = y(x, \theta^*)$, классифицируем неизвестные объекты (decision making)

# Как выбрать параметры модели?

Решить задачу оптимизации, чтобы получить значения $\theta^*$

Варианты подходов

- $\theta$ – фиксировано, но неизвестно: ищем $\theta$, согласующееся с обучающей выборкой
- $\theta$ – случайная величина, распределенная по известному закону: ищем параметры распределения

Обобщенные линейные модели

Метод максимального правдоподобия

Байесовский вывод

# 8 марта (is coming)

Setosa

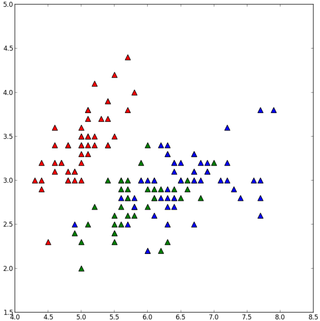
Versicolor

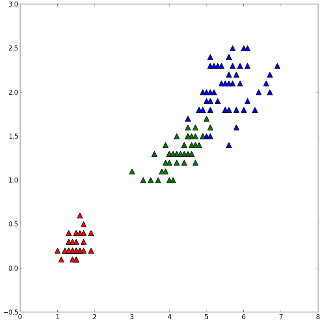
Virginica

## Задача

Определить вид ириса на основании длины чашелистика, ширины чашелистика, длины лепестка и ширины лепестка.

# Ирисы Фишера



# Линейные модели

Рассматривается случай 2 классов

Функция принятия решения

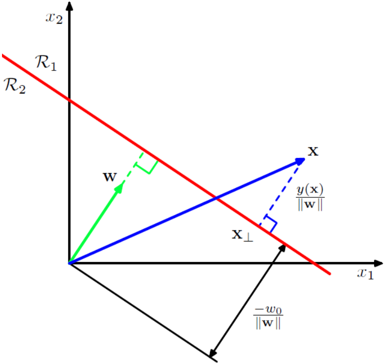
$$y(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + w_0$$

Регионы принятия решения

$$R_1 = \{\mathbf{x} \,:\, y(\mathbf{x}) > 0\}$$

$$R_2 = \{\mathbf{x} \,:\, y(\mathbf{x}) < 0\}$$

Задача
найти параметры модели $\mathbf{w}$, $w_0$



# Линейные модели: наблюдения

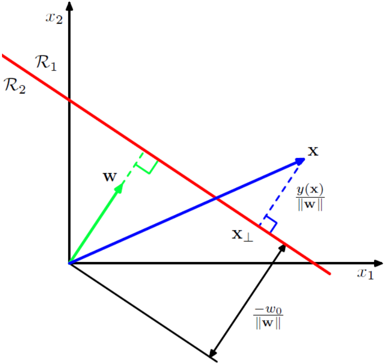
Разделяющая поверхность

$$\mathcal{D} = \{\mathbf{x} : \mathbf{w}^\top \mathbf{x} + w_0 = 0\}$$

1. $\mathbf{w}$ – нормаль к $\mathcal{D}$
2. $d = -\frac{w_0}{\|\mathbf{w}\|}$ – расстояние от центра координат до $\mathcal{D}$
3. $r(\mathbf{x}) = \frac{y(x)}{\|\mathbf{w}\|}$ – расстояние от $\mathcal{D}$ до $\mathbf{x}$

Положим $x_0 \equiv 1$, получим модель

$$y(\tilde{\mathbf{x}}) = \tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}$$



# Обобщенные линейные модели

Линейная модель

$$y(\mathbf{x}) = w_0 + \sum w_i x_i$$

Квадратичная модель

$$y(\mathbf{x}) = w_0 + \sum w_i x_i + \sum \sum w_{ij} x_i x_j$$

Обобщенная линейная модель

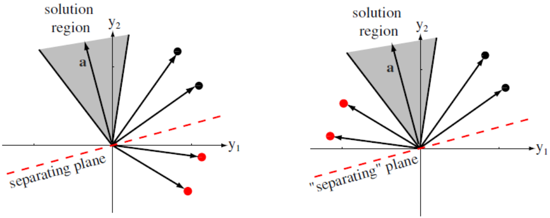
$$g(\mathbf{x}) = \sum a_i \phi_i(\mathbf{x}) = \mathbf{a}^\top \mathbf{y}$$

# Случай линейно разделимых классов

Обобщенная линейная модель

$$g(\mathbf{x}) = \sum a_i \phi_i(\mathbf{x}) = \mathbf{a}^\top \mathbf{y}$$

Дана обучающая выборка $\mathbf{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_N\}$




## Идея

Преобразовать объекты второго класса в обратные им и решать задачу оптимизации в области $a^T \mathbf{y}_i > 0, \ \forall i$

# Задача оптимизации

## Задача

Минимизируем критерий $J(a)$ при условиях $a^T \mathbf{y}_i > 0,\ \forall i$

Пусть $\mathcal{Y}$ – множество неправильно проклассифицированных объектов

- $J_e(a) = \sum_{\mathbf{y} \in \mathcal{Y}} 1$
- $J_p(a) = \sum_{\mathbf{y} \in \mathcal{Y}} -a^\top \mathbf{y}$
- $J_q(a) = \sum_{\mathbf{y} \in \mathcal{Y}} (a^\top \mathbf{y})^2$
- $J_r(a) = \sum_{\mathbf{y} \in \mathcal{Y}} \frac{(a^\top \mathbf{y})^2 - b}{\|\mathbf{y}\|}$

Улучшение: добавить отступы

# Градиентный спуск

1. initialise $a$, $J(a)$, $\eta(k)$, $\epsilon$, $k = 0$
2.   do $k \leftarrow k + 1$
3.     $a \leftarrow a - \eta(k) \nabla J(a)$
4.   until $\eta(k) \nabla J(a) < \epsilon$
5. return a
5. end

# Инкрементальный алгоритм

Рассматриваем $J_r(a) = \sum_{y \in \mathcal{Y}} \frac{(a^\top \mathbf{y})^2 - b}{\|y\|}$

1. initialise $a$, $\eta(k)$, $k = 0$
2. do $k \leftarrow k + 1$
3. if $\mathbf{y}_k$ is misclassified $a \leftarrow a - \eta(k) \frac{(a^\top \mathbf{y_k})^2 - b}{\|\mathbf{y}_k\|^2} \mathbf{y}_k$
4. until no errors left
5. return a
6. end

# Случай линейно неразделимых классов

- Использовать $\eta(k) \to 0$ при $k \to \infty$
- От системы неравенств перейти к системе линейных уравнений
- Линейное программирование

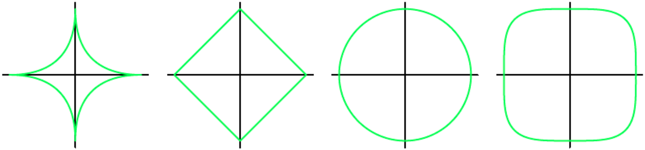
# Снова переобучение

Оптимизируем критерий с регуляризацией

$$J_1(a) = J(a) + \lambda J_R(a)$$

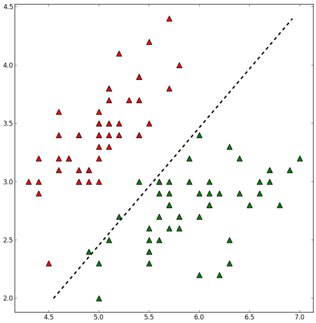
$\lambda$ – коэффициент регуляризации

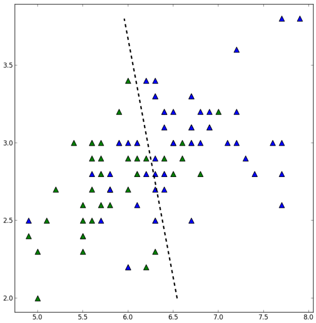
$$J_R(a) = \sum |a_j|^q$$



$q = 0.5$ $q = 1$ $q = 2$ $q = 4$

# Перцептрон: результаты




ТЕХНОСФЕРА @mail.ru

# Метод максимального правдоподобия

## Задача

Дана обучающая выборка $\mathbf{X}$. Предполагая, что распределение $p(\mathbf{x}|\theta)$ известно, найти значения параметров $\theta$.

Интуиция: найти такие $\theta$, которые максимизируют вероятность $P(\mathbf{X}|\theta)$.

При предположении, что обучающие образцы независимы, имеем

$$P(\mathbf{X}|\theta) = \prod p(\mathbf{x}_i|\theta)$$

Функция правдоподобия

$$l(\theta) = \log P(\mathbf{X}|\theta) = \sum \log p(\mathbf{x}_i|\theta)$$

Требуется найти

$$\theta = \arg\max_{\theta} l(\theta)$$

# Нормальное распределение

$$p(x|\mu) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

## Задача

Дана выборка **X** объектов $x$, распределенных согласно одномерному нормальному закону $N(\mu, \sigma^2)$. Используя принцип максимального правдоподобия, оценить значение $\mu$ при известном значении $\sigma$.

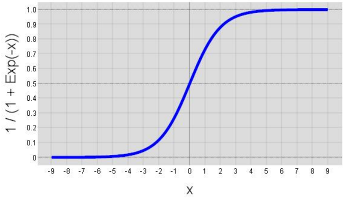
# Вероятностная линейная модель

Рассматриваем 2 класса

$$p(C_1|x) = \frac{p(x|C_1)p(C_1)}{p(x|C_1)p(C_1) + p(x|C_2)p(C_2)} = \frac{1}{1+e^{-a}} = \sigma(a)$$

$$a = \ln \frac{p(x|C_1)p(C_1)}{p(x|C_2)p(C_2)}$$

$\sigma(a)$ – сигмоид-функция, $a = \ln(\sigma/(1-\sigma))$




Упражнение: $p(\mathbf{x}|C_k) = \mathcal{N}(\mu_k, \mathbf{\Sigma})$. Проверить, что $p(C_k|x) = \sigma(\mathbf{w}^\top \mathbf{x} + w_0)$

# (Еще более) обобщенная линеная модель

Базисные функции $\phi_n(\mathbf{x})$

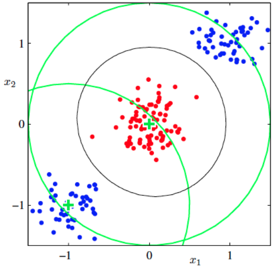
$$\phi_n(\mathbf{x}) = \exp\left[-\frac{(x-\mu_n)^2}{2s^2}\right]$$

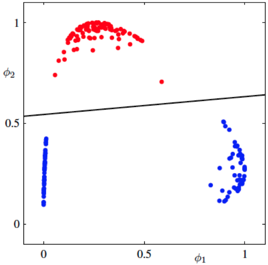
Функция активации $f(a)$

$$f(a) = \sigma(a)$$

(Совсем) обобщенная линейная модель

$$y(\mathbf{x}, \mathbf{w}) = f(\mathbf{w}^\top \phi(\mathbf{x}))$$



# Логистическая регрессия

дано

$$\{\phi_n = \phi(\mathbf{x}_n), t_n\},\ \ t_n \in \{0, 1\},\ \ n = 1 \dots N$$

модель

$$p(C_1|\phi) = y(\phi) = \sigma(\mathbf{w}^\top \phi)$$

функция правдоподобия

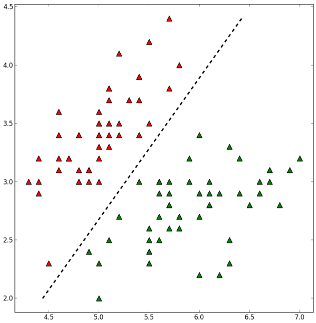
$$l(\mathbf{w}) = \log\left[\prod_{n=1}^{N} p^{t_n}(C_1|\phi_n)(1 - p(C_1|\phi_n))^{1-t_n}\right] =$$

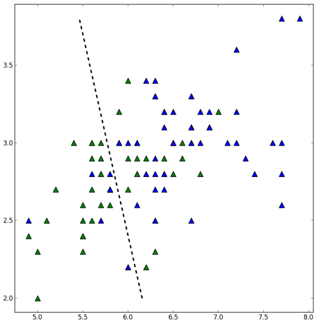
$$= \sum_{n=1}^{N} t_n \log p(C_1|\phi_n) + (1 - t_n)\log(1 - p(C_1|\phi_n)) = -J_e(\mathbf{w})$$

градиент

$$\nabla J_e(\mathbf{w}) = \sum_{n=1}^{N}(p(C_1|\phi_n) - t_n)\phi_n \to min_{\mathbf{w}}$$

# Логистическая регрессия: результаты



# Байесовский вывод

## Дано

- плотность вероятности $p(\mathbf{x}|\theta)$
- априорная плотность $p(\theta)$
- выборка $\mathbf{X} = \{x_1, \dots, x_N\}$

## Найти

апостериорную плотность $p(\theta|\mathbf{X})$

$$p(\mathbf{x}|\mathbf{X}) = \int p(\mathbf{x}|\theta)p(\theta|\mathbf{X})d\theta$$

$$p(\theta|\mathbf{X}) = \frac{p(\mathbf{X}|\theta)p(\theta)}{\int p(\mathbf{X}|\theta)p(\theta)d\theta}$$

$$p(\mathbf{X}|\theta) = \prod_n p(x_n|\theta)$$

# Мультиклассовая классификация

## Задача

Использовать бинарный линейный классификатор для мультиклассовой классификации Ирисов. Какие идеи?

Интерфейс классификатора `clf`

- Использовать выборку **x**, **y** для обучения
  `clf.fit(x, y)`
- Предсказать класс объектов в **x**
  `y = clf.predict(x)`
- Предсказать вероятности классов для **x**
  `y = clf.predict_proba(x)`
- Вычислить значение функции решения в $x$
  `d = clf.decision_function(x)`

# Домашнее задание 1

## Линейные модели

Реализовать на выбор

- Линейная классификация методом градиентного спуска
- Линейная регрессия методом градиентного спуска
- Линейная классификация инкрементальным методом
- Линейная регрессия инкрементальным методом

Ключевые даты

- До 2014/03/14 23.59 выбрать задачу и ответственного в группе
- До 2014/03/21 00.00 предоставить решение задания

# Спасибо!

Обратная связь