Логистическая регрессия

Машинное обучение — осень 2015

Количественная аналитика

Содержание

- теоретические основы метода
- пример практической реализации в «R»
- домашнее задание

Основные обозначения

$\vec{y}_{[m\times 1]}$ — вектор значений объясняемой переменной

$y^{(i)} \in \{0; 1\},\ i \in \{1; \dots; m\}$

$X_{[m\times(n+1)]}$ — матрица значений объясняющих переменных

$\vec{x}^{(i)} = \left(x_0^{(i)}, \dots, x_n^{(i)}\right)^T \in R^{n+1},\ i \in \{1; \dots; m\}$

$x_0^{(i)} = 1,\ i \in \{1; \dots; m\}$

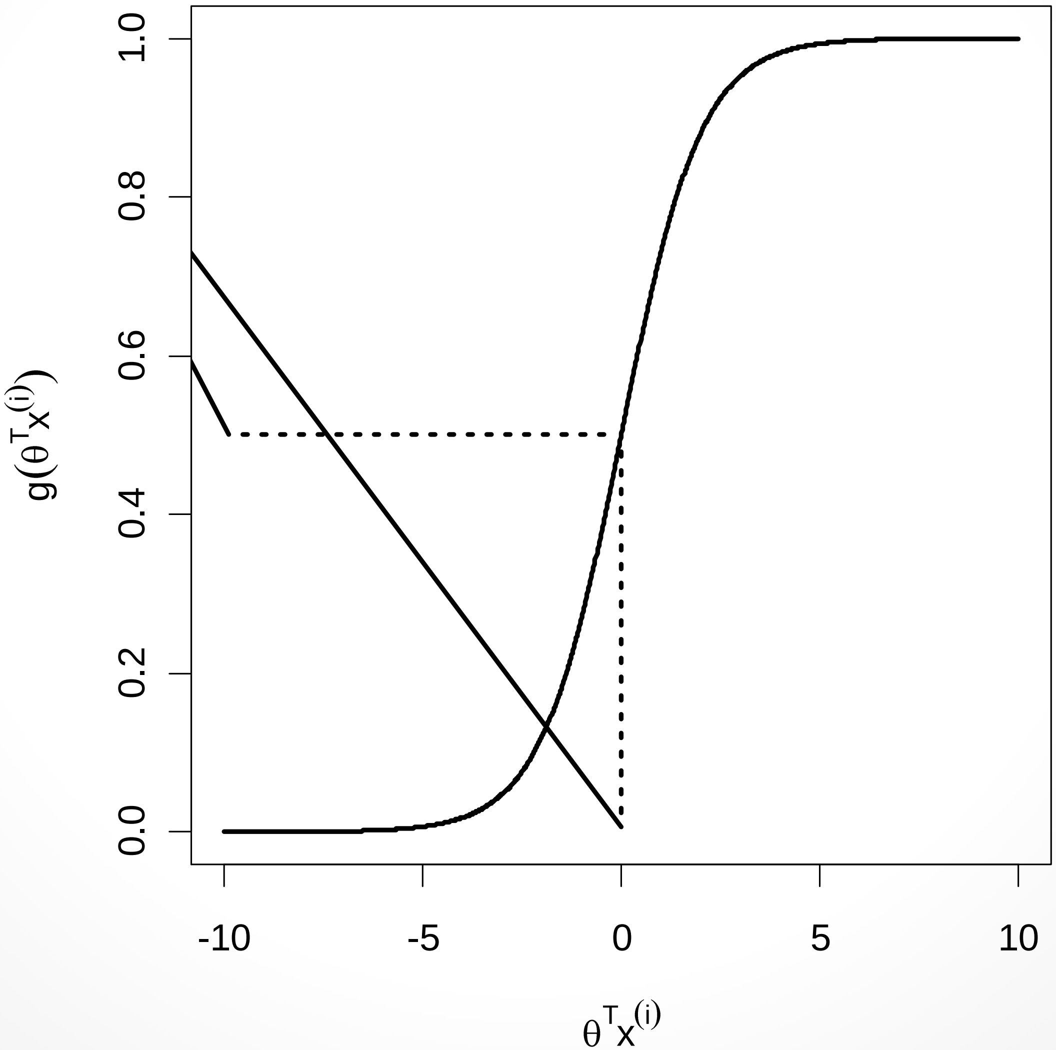
$\hat{y}^{(i)} = h_\theta\left(\vec{x}^{(i)}\right) = \frac{1}{1+\exp\left(-\vec{\theta}^T \vec{x}^{(i)}\right)}$ — гипотеза

$\vec{\theta}_{[n\times 1]}$ — вектор параметров

$h_\theta\left(\vec{x}^{(i)}\right)$ интерпретируется как $P\left(y^{(i)} = 1\right)$

Логистическая функция $g(z)$

$$h_\theta\left(\vec{x}^{(i)}\right) = g\left(\vec{\theta}^T \vec{x}^{(i)}\right),\ g(z) = \frac{1}{1+\exp(-z)}$$



Граница принятия решения

Формируем прогноз $\hat{y}^{(i)} = 1$, если $h_\theta\left(\vec{x}^{(i)}\right) \geq 0.5 \Rightarrow \vec{\theta}^T \vec{x}^{(i)} \geq 0$

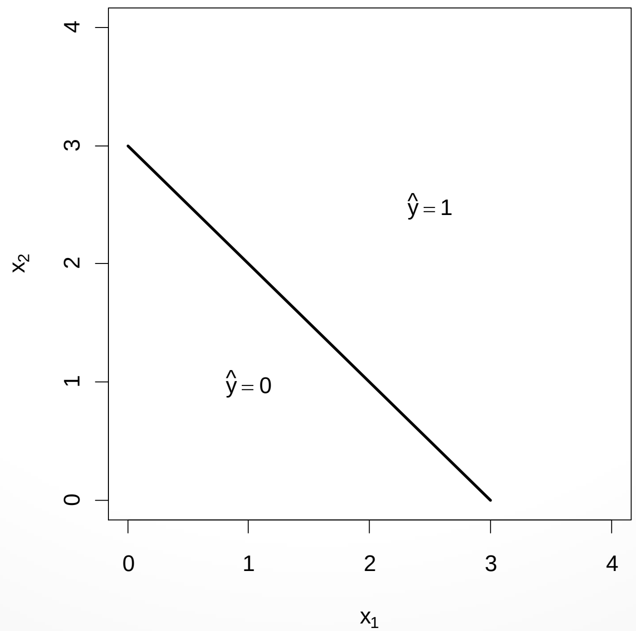
$\hat{y}^{(i)} = 0$, если $h_\theta\left(\vec{x}^{(i)}\right) < 0.5 \Rightarrow \vec{\theta}^T \vec{x}^{(i)} < 0$

Пусть $n = 2$, тогда $h_\theta(\vec{x}) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$

Пусть $\vec{\theta} = (-3,1,1)^T$, тогда

$\hat{y} = 1$, если $\vec{\theta}^T \vec{x} = -3 + x_1 + x_2 \geq 0 \Rightarrow x_1 + x_2 \geq 3$

Прямая $x_1 + x_2 = 3$ называется границей принятия решения



Функция потерь

$J(\vec{\theta}) = \frac{1}{m}\sum_{i=1}^{m} cost\left(h_\theta\left(\vec{x}^{(i)}\right), y^{(i)}\right)$

$cost(\dots)$ — потери при классификации *i*-го наблюдения

$$cost\left(h_\theta\left(\vec{x}^{(i)}\right), y^{(i)}\right) = \begin{cases} -\log h_\theta\left(\vec{x}^{(i)}\right), & y^{(i)} = 1 \\ -\log\left(1 - h_\theta\left(\vec{x}^{(i)}\right)\right), & y^{(i)} = 0 \end{cases}$$

Таким образом,

$$J(\vec{\theta}) = -\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)}\log h_\theta\left(\vec{x}^{(i)}\right) + \left(1 - y^{(i)}\right)\log\left(1 - h_\theta\left(\vec{x}^{(i)}\right)\right)\right)$$

$$J(\vec{\theta}) \to \min_{\vec{\theta}}$$

Векторизованная форма:

$$J(\vec{\theta}) = -\frac{1}{m}\vec{y}^T\log\left(g\left(X\vec{\theta}\right)\right) - \frac{1}{m}(\vec{\imath}_m - \vec{y})^T\log\left(\vec{\imath}_m - g\left(X\vec{\theta}\right)\right)$$

$\vec{\imath}_m$ — вектор единиц длиной *m*

Градиент функции потерь

Некоторые оптимизационные процедуры (ньютоновские и квази-ньютоновские) работают быстрее, если известен градиент минимизируемой функции

$$\frac{\partial}{\partial\theta_j}J(\vec{\theta}) = \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(\vec{x}^{(i)}\right) - y^{(i)}\right)x_j^{(i)}$$

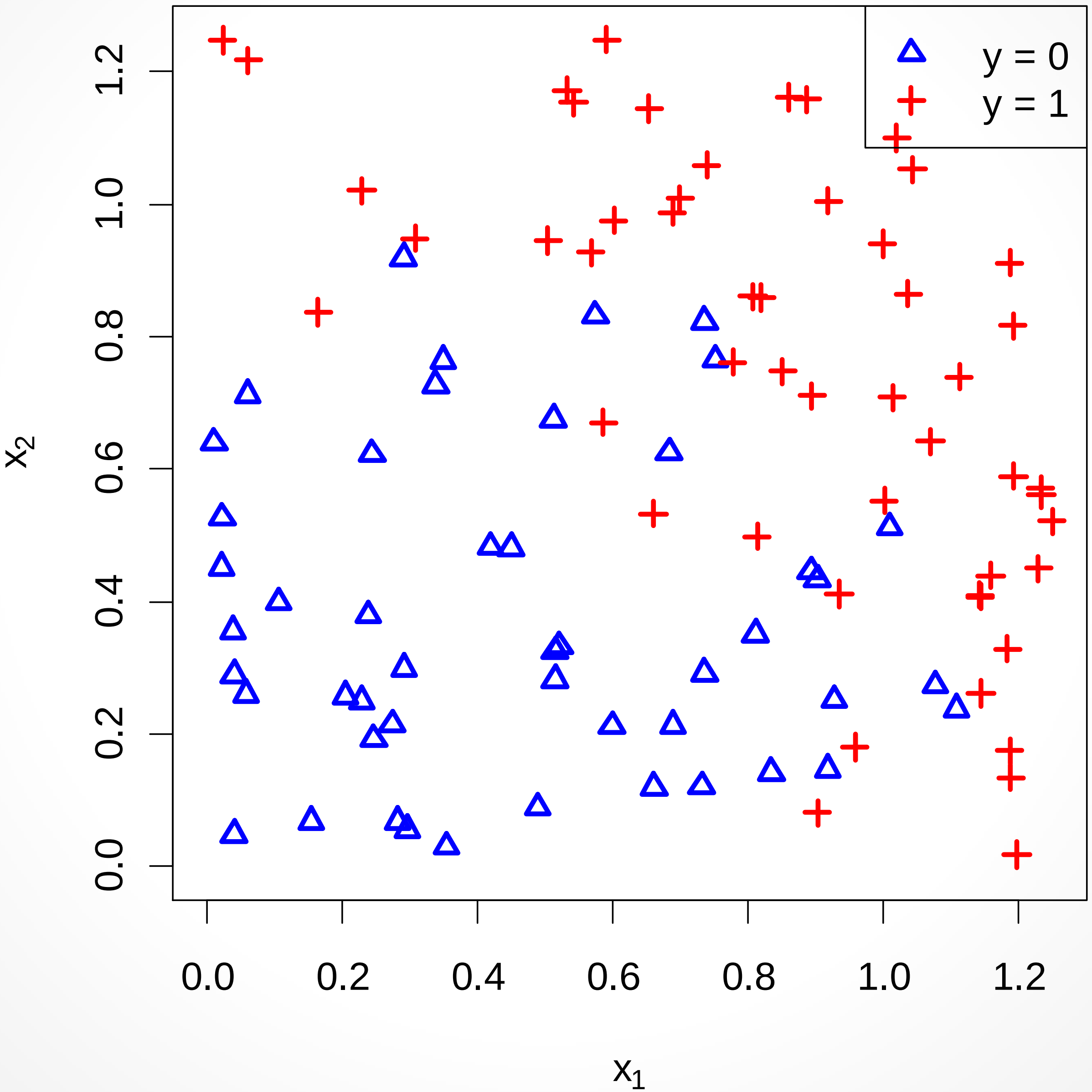
Векторизованная форма:

$\nabla J(\vec{\theta}) = \frac{1}{m}X^T\left(g\left(X\vec{\theta}\right) - \vec{y}\right),$ где

$\nabla J(\vec{\theta})$ — вектор производных функции $J(\vec{\theta})$

Логистическая регрессия в R

Визуализация исходных данных



Функция потерь и градиент

\# исходные данные и их форматы

```
y <- cbind(y); X <- as.matrix(X)
X <- cbind(1, X) # если в матрице X нет единичного столбца
m <- nrow(X); n <- ncol(X) - 1
```

\# логистическая функция

```
g <- function(z) 1/(1+exp(-z))


J <- function(theta) {
  m <- nrow(X)
  # вычисление гипотезы hθ(x⃗)
  # theta и y — векторы-столбцы, X — матрица
  h.theta <- g(X%*%theta)
  -t(y)%*%log(h.theta)/m - t(1-y)%*%log(1-h.theta)/m
}

gradJ <- function(theta) {
  m <- nrow(X)
  t(X)%*%(g(X%*%theta)-y)/m
}
```

Подгонка параметров $\vec{\theta}$

\# начальные значения

```
theta0 <- cbind(rep(0,times=n+1))
```

\# численная оптимизация

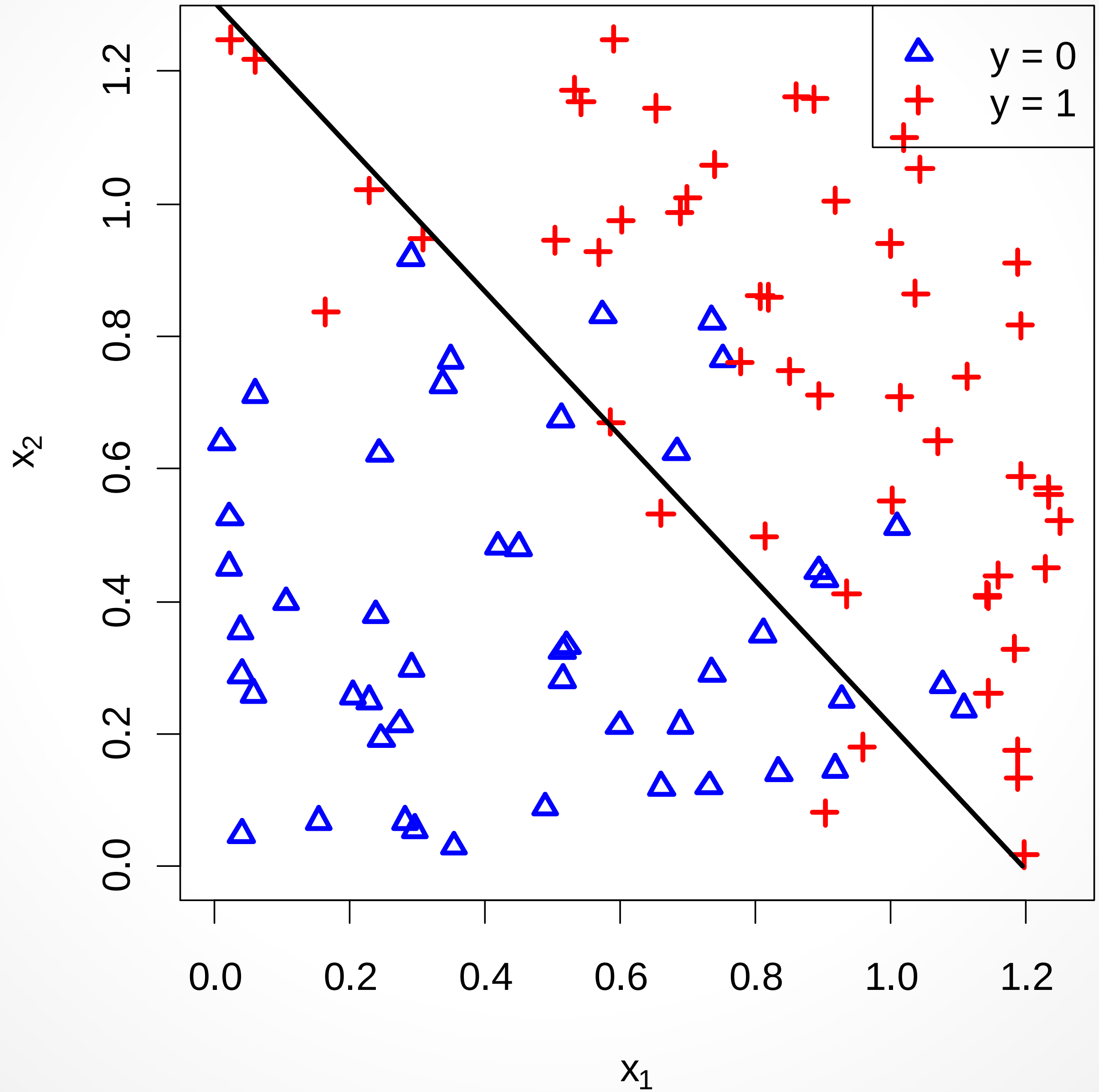
```
opt <- optim(fn=J, gr=gradJ, par=theta0, method="BFGS")
theta <- opt$par; Jval <- opt$value

list(theta=as.vector(theta),J=Jval)
$theta                                    $J
[1] -8.143582  6.248701  6.815362    [1] 0.3062095
```

\# визуализация линейной границы принятия решения

\# по двум точкам прямой $\theta_0 + \theta_1 x_1 + \theta_2 x_2 = 0$

```
x1 <- c(0,-theta[1]/theta[3])
x2 <- c(-theta[1]/theta[2],0)
lines(x1,x2,type="l",lwd=3)
```

Визуализация линейной границы принятия решений (ГПР)



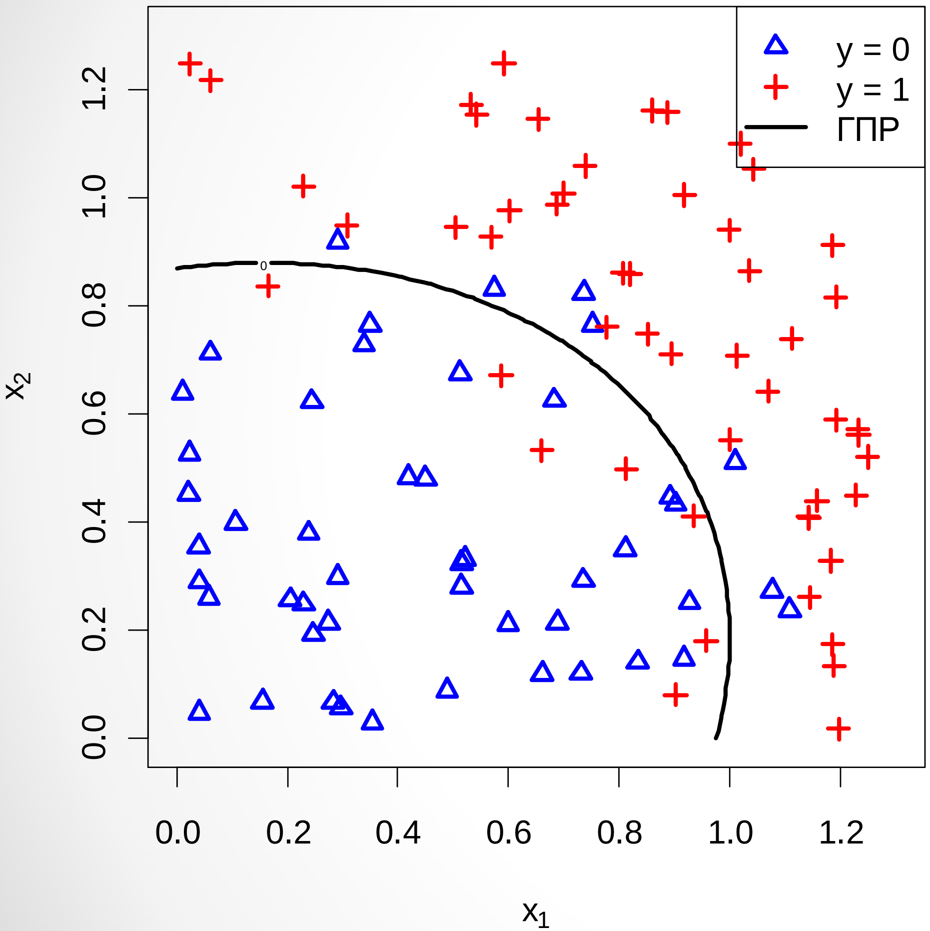
Избыточная подгонка модели

Повышение качества подгонки моделей достигается добавлением в неё новых объясняющих переменных или использованием полиномиальных версий существующих

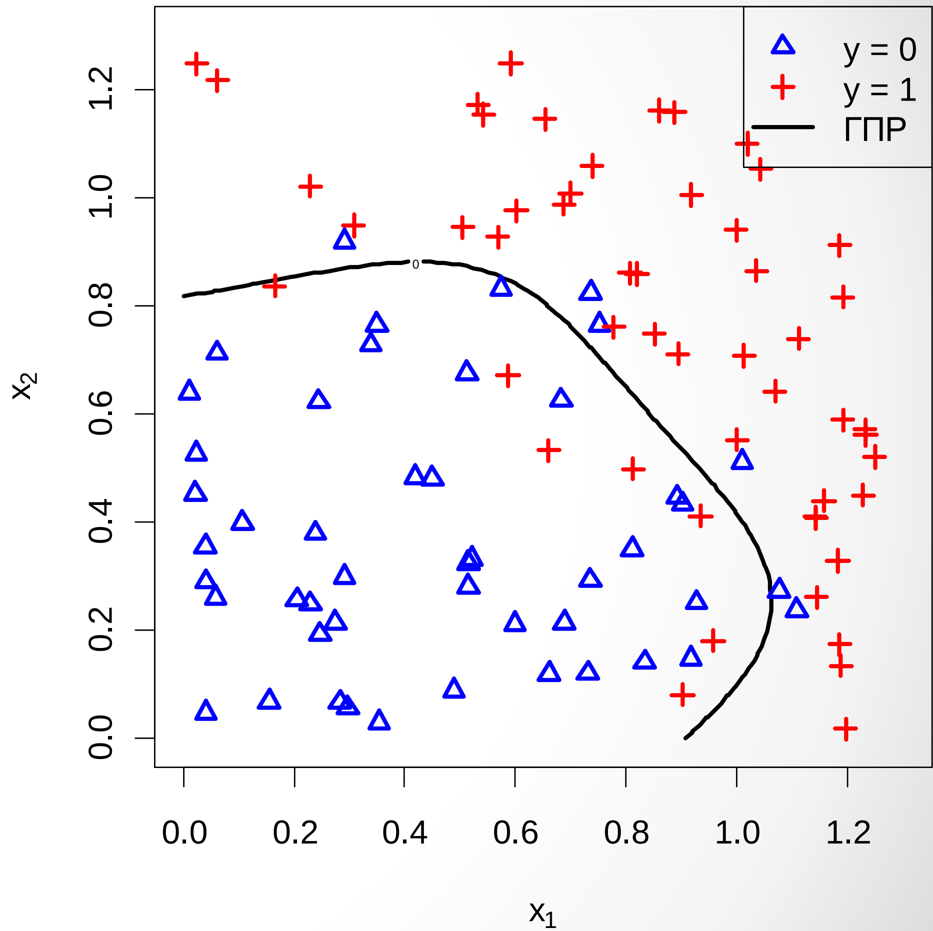
Однако это может привести к эффекту «избыточной подгонки» модели (overfitting), когда ошибка на обучающей выборки низка, а на тестовой — очень высока

Полиномиальная ГПР

$d = 2$



$d = 6$



Методы устранения избыточной подгонки

1. Сократить набор объясняющих переменных / уменьшить порядок полинома
   - вручную
   - с помощью алгоритма выбора модели (model selection algorithm)

2. Регуляризация
   - уменьшение значений параметров $\vec{\theta}$

Регуляризованные функция потерь и градиент

Регуляризованная функция потерь

$$J(\vec{\theta}) = -\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)}\log h_\theta\left(\vec{x}^{(i)}\right) + \left(1 - y^{(i)}\right)\log(1 - h_\theta\left(\vec{x}^{(i)}\right))\right.$$
$$+ \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$$

λ — параметр регуляризации

Регуляризованный градиент

$$\begin{cases} \dfrac{\partial}{\partial\theta_0}J(\vec{\theta}) = \dfrac{1}{m}\displaystyle\sum_{i=1}^{m}\left(h_\theta\left(\vec{x}^{(i)}\right) - y^{(i)}\right)x_0^{(i)}, & j = 0 \\ \dfrac{\partial}{\partial\theta_j}J(\vec{\theta}) = \dfrac{1}{m}\displaystyle\sum_{i=1}^{m}\left(h_\theta\left(\vec{x}^{(i)}\right) - y^{(i)}\right)x_j^{(i)} + \dfrac{\lambda}{m}\theta_j, & j \neq 0 \end{cases}$$

Регуляризация в R

\# функция потерь

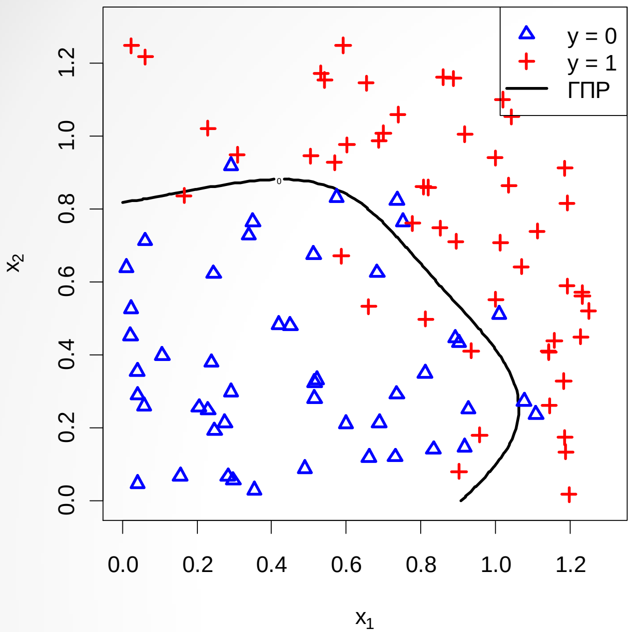
```
J.reg <- function(theta) {
  m <- nrow(X)
  h.theta <- g(X%*%theta)
  # нерегуляризованная функция
  J <- -t(y)%*%log(h.theta)/m - t(1-y)%*%log(1-h.theta)/m
  # регулязационная составляющая
  reg <- lambda*sum(theta^2)/(2*m)
  J + reg
}
```

\# градиент

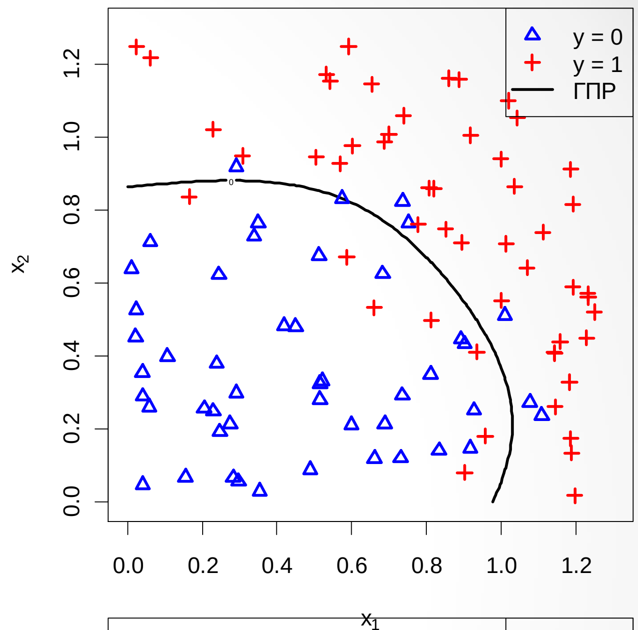
```
gradJ.reg <- function(theta) {
  m <- nrow(X)
  # нерегуляризованный градиент
  grad <- t(X)%*%(g(X%*%theta)-y)/m
  # регулязационная составляющая
  reg <- lambda/m * theta; reg[1] <- 0
  grad + reg
}
```

Влияние параметра λ

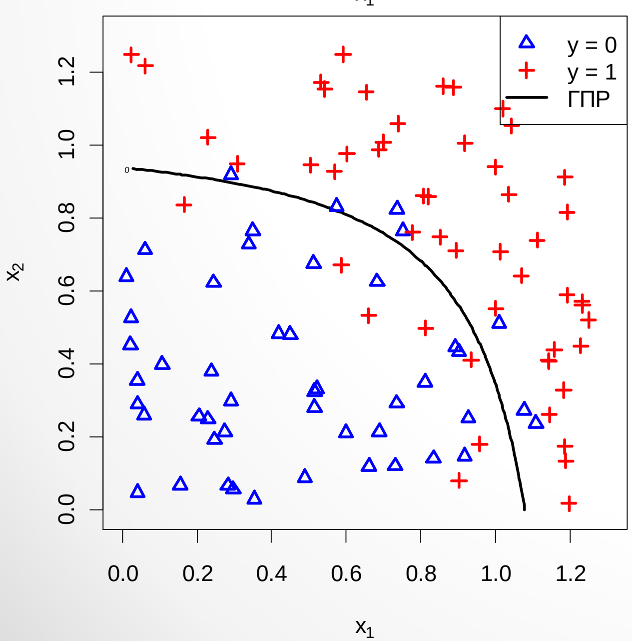
$\lambda = 0$



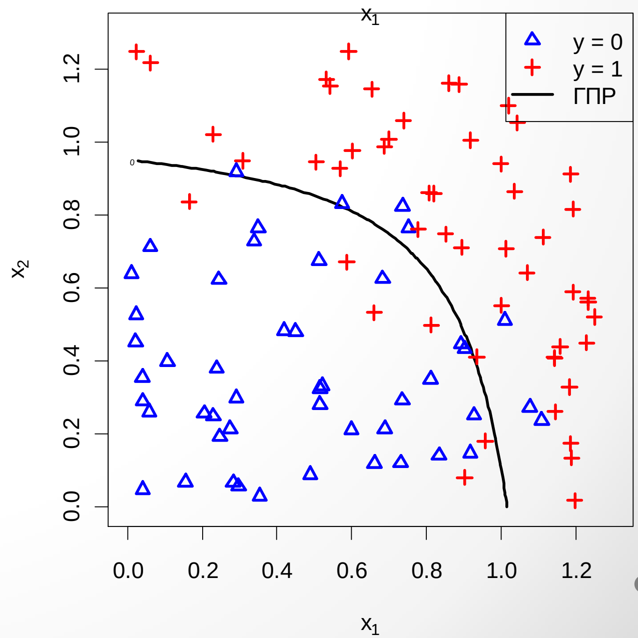
$\lambda = 0{,}01$



$\lambda = 1$



$\lambda = 10$



Домашнее задание

TBA