RVM

Атанов Андрей

HSE

3 Октября 2016

Немного про RVM

Немного про RVM

Данные

$X = \{x_i\}_{i=1}^n, x_i \in \mathbb{R}^d$ - объекты выборки и $t \in \mathbb{R}^n$ - вектор целевых переменных.

Немного про RVM

Данные

$X = \{x_i\}_{i=1}^n, x_i \in \mathbb{R}^d$ - объекты выборки и $t \in \mathbb{R}^n$ - вектор целевых переменных.

Посыл

$$t_n \sim \mathcal{N}(y(w, x_i), \sigma^2)$$

Немного про RVM

Данные

$X = \{x_i\}_{i=1}^n, x_i \in \mathbb{R}^d$ - объекты выборки и $t \in \mathbb{R}^n$ - вектор целевых переменных.

Посыл

$$t_n \sim \mathcal{N}(y(w, x_i), \sigma^2)$$

$$y(w, x) = \sum_{i=1}^{D} w_i \phi_i(x) = w^T \phi(x)$$

$\phi(x) = (\phi_1(x), ..., \phi_D(x))$ - вектор признаков.

Немного про RVM

Функция правдоподобия

$$\mathrm{p}(\mathrm{t}|\mathrm{X}, \mathrm{w}, \sigma^2) = (2\pi\sigma^2)^{\mathrm{n}/2} \exp\left\{-\frac{1}{2\sigma^2}||\Phi\mathrm{w} - \mathrm{t}||^2\right\} \Rightarrow \mathrm{w}_{\mathrm{ML}}$$

Немного про RVM

Функция правдоподобия

$$\mathrm{p}(\mathrm{t}|\mathrm{X}, \mathrm{w}, \sigma^2) = (2\pi\sigma^2)^{\mathrm{n}/2} \exp\left\{-\frac{1}{2\sigma^2}||\Phi\mathrm{w} - \mathrm{t}||^2\right\} \Rightarrow \mathrm{w}_{\mathrm{ML}}$$

Есть серьезная проблема - переобучение.

Немного про RVM

Функция правдоподобия

$$\mathrm{p}(\mathrm{t}|\mathrm{X},\mathrm{w},\sigma^2) = (2\pi\sigma^2)^{\mathrm{n}/2}\exp\left\{-\frac{1}{2\sigma^2}||\Phi\mathrm{w}-\mathrm{t}||^2\right\} \Rightarrow \mathrm{w}_{\mathrm{ML}}$$

Есть серьезная проблема - переобучение.

Регуляризация

$$\mathrm{w} \sim \mathcal{N}(0, \mathrm{A}^{-1})$$

$\mathrm{A} = \mathrm{diag}(\alpha_1, ..., \alpha_{\mathrm{D}})$

Немного про RVM

Функция правдоподобия

$$\mathrm{p}(\mathrm{t}|\mathrm{X},\mathrm{w},\sigma^2) = (2\pi\sigma^2)^{\mathrm{n}/2}\exp\left\{-\frac{1}{2\sigma^2}||\Phi\mathrm{w}-\mathrm{t}||^2\right\} \Rightarrow \mathrm{w}_{\mathrm{ML}}$$

Есть серьезная проблема - переобучение.

Регуляризация

$$\mathrm{w} \sim \mathcal{N}(0, \mathrm{A}^{-1})$$

$\mathrm{A} = \mathrm{diag}(\alpha_1, ..., \alpha_{\mathrm{D}})$

Апостериорное распределение

$$\mathrm{p}(\mathrm{w}|\mathrm{t}) \propto \mathrm{p}(\mathrm{t}|\mathrm{w})\mathrm{p}(\mathrm{w}|\alpha) \Rightarrow \mathrm{w}_{\mathrm{MAP}}$$

Немного про RVM

Хотим

Зная обучающую выборку $\{\mathrm{X}, \mathrm{t}\}$, оценить для нового объекта x^* целевую переменную t^* (восстановить плотность).

$$\mathrm{p}(\mathrm{t}^*|\mathrm{t}) = \int \mathrm{p}(\mathrm{t}^*|\mathrm{w}, \alpha, \sigma^2)\mathrm{p}(\alpha, \sigma^2, \mathrm{w}|\mathrm{t})\mathrm{dwd}\alpha\mathrm{d}\sigma^2$$

Немного про RVM

Хотим

Зная обучающую выборку $\{X, t\}$, оценить для нового объекта x^* целевую переменную t^* (восстановить плотность).

$$p(t^*|t) = \int p(t^*|w, \alpha, \sigma^2)p(\alpha, \sigma^2, w|t)dwd\alpha d\sigma^2$$

Проблема

$$p(w, \alpha, \sigma^2|t) = \frac{p(t|\sigma^2, \alpha, w)p(\alpha, \sigma^2, w)}{p(t)}$$

Немного про RVM

Хотим

Зная обучающую выборку $\{X, t\}$, оценить для нового объекта x^* целевую переменную t^* (восстановить плотность).

$$p(t^*|t) = \int p(t^*|w, \alpha, \sigma^2)p(\alpha, \sigma^2, w|t)dwd\alpha d\sigma^2$$

Проблема

$$p(w, \alpha, \sigma^2|t) = \frac{p(t|\sigma^2, \alpha, w)p(\alpha, \sigma^2, w)}{p(t)}$$

$p(t)$ - нельзя вычислить аналитически.

Немного про RVM

Хотим

Зная обучающую выборку $\{X, t\}$, оценить для нового объекта x^* целевую переменную t^* (восстановить плотность).

$$p(t^*|t) = \int p(t^*|w, \alpha, \sigma^2)p(\alpha, \sigma^2, w|t)dwd\alpha d\sigma^2$$

Проблема

$$p(w, \alpha, \sigma^2|t) = \frac{p(t|\sigma^2, \alpha, w)p(\alpha, \sigma^2, w)}{p(t)}$$

$p(t)$ - нельзя вычислить аналитически.

Решение

$$p(w, \alpha, \sigma^2|t) = p(w|t, \alpha, \sigma^2)p(\alpha, \sigma^2|t)$$

Выбор модели (настройка гиперпараметров)

Максимизация обоснованности

$$\mathrm{p}(\alpha, \sigma^2|\mathrm{t}) \propto \mathrm{p}(\mathrm{t}|\alpha, \sigma^2)\mathrm{p}(\alpha)\mathrm{p}(\sigma^2)$$

$$\alpha_{\mathrm{MP}}, \sigma^2_{\mathrm{MP}} = \mathrm{argmax}\{\mathrm{p}(\mathrm{t}|\alpha, \sigma^2)\}$$

Для задачи регрессии обоснованность можно посчитать аналитически, для классификации приходится пользоваться аппроксимацией Лапласа.

Выбор модели (настройка гиперпараметров)

Максимизация обоснованности

$$\mathrm{p}(\alpha, \sigma^2|\mathrm{t}) \propto \mathrm{p}(\mathrm{t}|\alpha, \sigma^2)\mathrm{p}(\alpha)\mathrm{p}(\sigma^2)$$

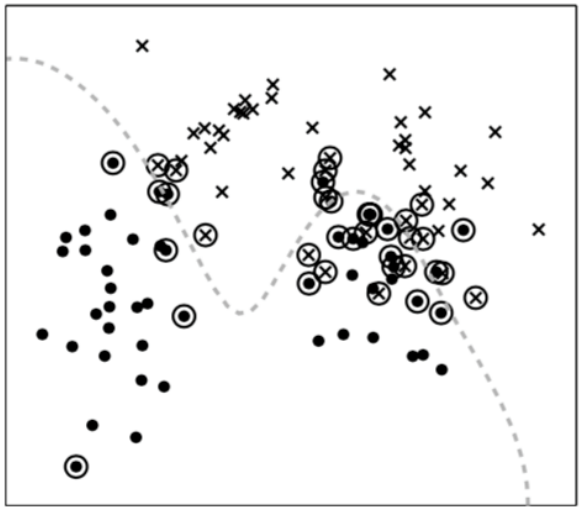
$$\alpha_{\mathrm{MP}}, \sigma^2_{\mathrm{MP}} = \mathrm{argmax}\{\mathrm{p}(\mathrm{t}|\alpha, \sigma^2)\}$$

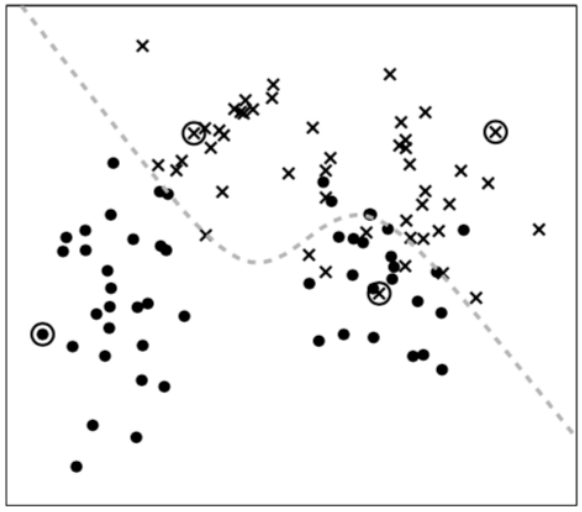
Для задачи регрессии обоснованность можно посчитать аналитически, для классификации приходится пользоваться аппроксимацией Лапласа.

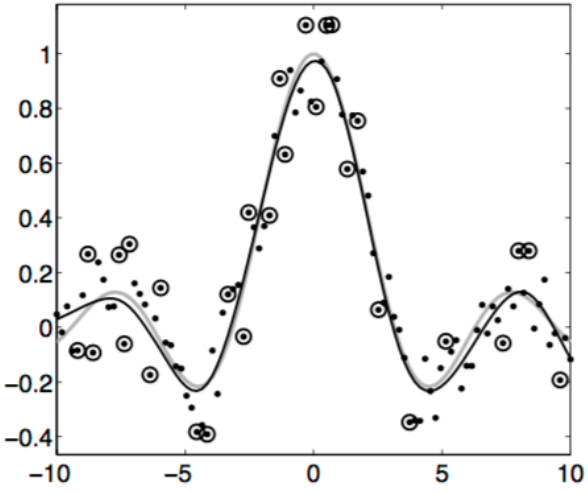
Итоговое распределение

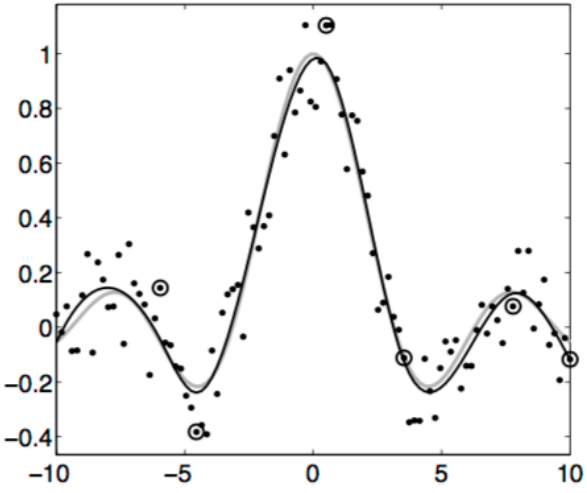
$$\begin{aligned}\mathrm{p}(\mathrm{t}^*|\mathrm{t}, \alpha_{\mathrm{MP}}, \sigma^2_{\mathrm{MP}}) &= \int \mathrm{p}(\mathrm{t}^*|\mathrm{w}, \alpha_{\mathrm{MP}}, \sigma^2_{\mathrm{MP}})\mathrm{p}(\mathrm{w}|\mathrm{t}, \alpha_{\mathrm{MP}}, \sigma^2_{\mathrm{MP}})\mathrm{dw} \\ &= \mathcal{N}(\mathrm{t}^*|\mathrm{y}_*, \sigma^2_*)\end{aligned}$$

Пример









RVM

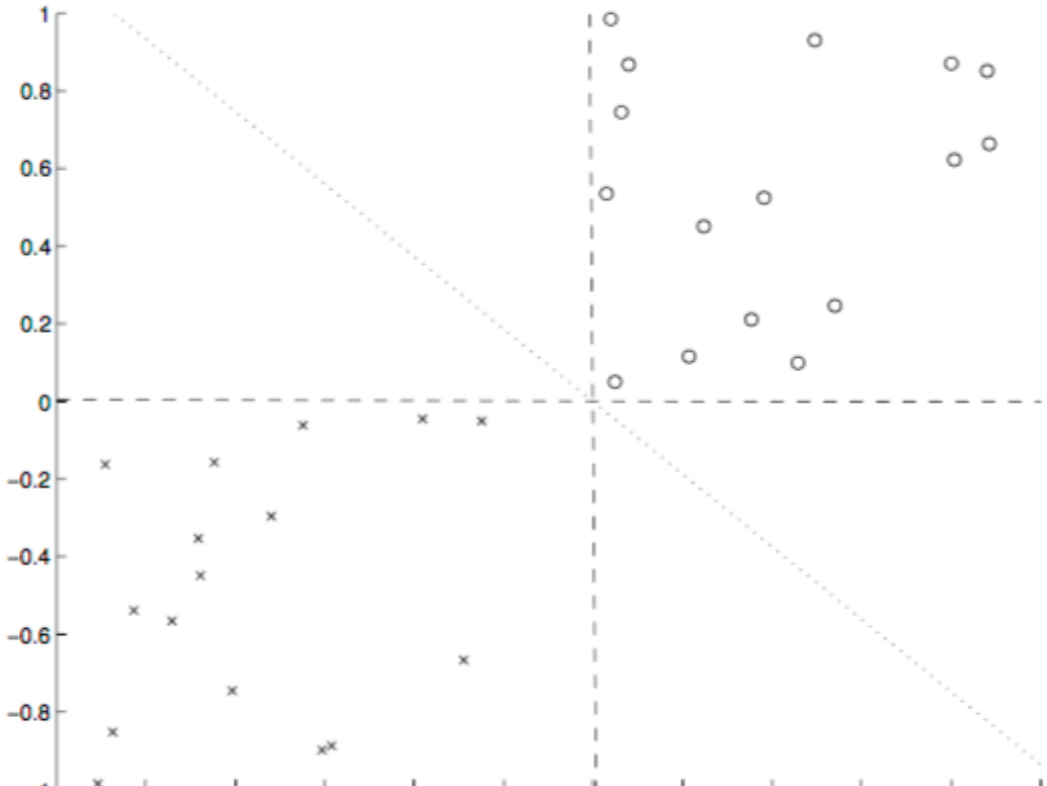
Обсуждение

- $\alpha_i \to +\infty \Rightarrow$ вес i-го признака обнуляется и соответствующий признак удаляется из модели. Обученная модель оказывается намного разреженнее по сравнению с SVM. Данный эффект носит название Auto Relevance Determination.

RVM

Обсуждение

- $\alpha_i \to +\infty \Rightarrow$ вес i-го признака обнуляется и соответствующий признак удаляется из модели. Обученная модель оказывается намного разреженнее по сравнению с SVM. Данный эффект носит название Auto Relevance Determination.



RVM

Обсуждение

- Автоматическая настройка параметров α, σ^2.

RVM

Обсуждение

- Автоматическая настройка параметров α, σ^2.
- Использование всей выборке для настройки параметров и гиперпараметров

RVM

Обсуждение

- Автоматическая настройка параметров α, σ^2.
- Использование всей выборке для настройки параметров и гиперпараметров
- НО все-таки необходимо подбирать параметры ядровой функции

Подбор параметра ядровой функции

Идея

При использовании ядровой ф-ии в качестве признаков, правдоподобие так же зависит и от параметра, входящего в нее. Например, для RBF ядра $\mathrm{K}(\mathrm{x_i}, \mathrm{x_j}) = \exp\{-\gamma||\mathrm{x_i} - \mathrm{x_j}||^2\}$:

$$\mathrm{p}(\mathrm{t}|\alpha, \sigma^2, \gamma) \to \max_{\gamma \in \mathrm{C}}$$

C - конечное множетсво.

Результаты

Кросс-валидация

Data set	Polynomial Kernel			RBF Kernel		
	CV R^2	Test R^2	$\log p(t\|\alpha, \sigma^2)$	CV R^2	Test R^2	$\log p(t\|\alpha, \sigma^2)$
Airfoil	0.509	0.591	-221	0.509	0.645	-198
Concrete	0.61	0.733	-169	0.672	0.778	-178
CCPP	0.927	0.926	-61	0.933	0.935	-42
Life expectancy	0.882	0.345	-43	0.922	0.814	-23
Friedman #3	0.482	0.734	-282	0.765	0.879	-196
Boston Housing	0.834	0.906	-308	0.862	0.919	-133

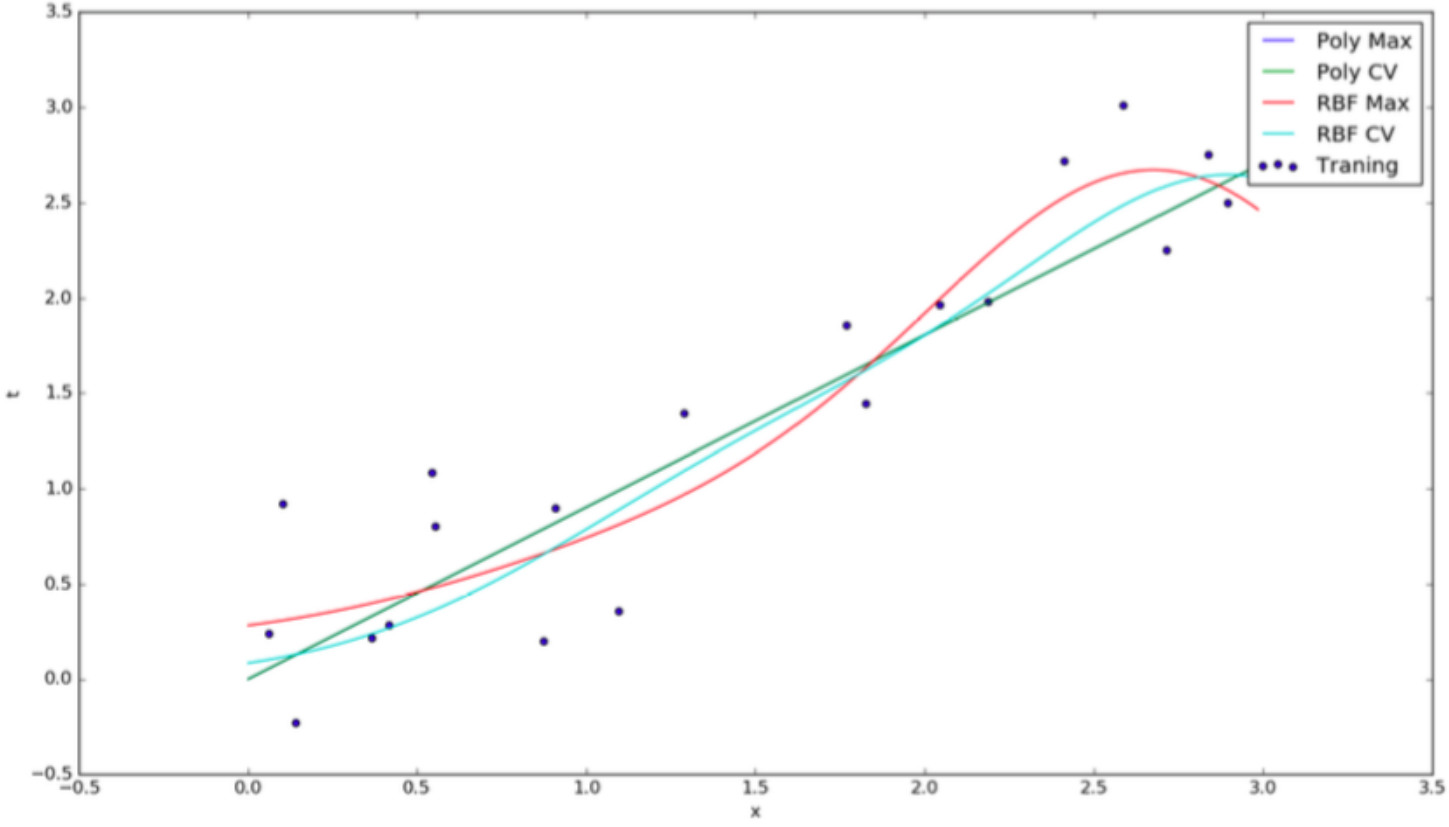
Максимизация обоснованности

Data set	Polynomial Kernel			RBF Kernel		
	CV R^2	Test R^2	$\log p(t\|\alpha, \sigma^2)$	CV R^2	Test R^2	$\log p(t\|\alpha, \sigma^2)$
Airfoil	0.495	0.591	-221	0.666	0.739	-168
Concrete	0.656	0.722	-152	0.663	0.762	-155
Poly data	0.964	0.985	21	0.969	0.991	18
CCPP	0.927	0.926	-58	0.933	0.935	-35
Life expectancy	0.724	0.328	-41	0.916	0.888	-21
Friedman #3	0.478	0.721	-269	0.049	0.095	-136
Boston Housing	0.789	0.925	-255	0.478	0.713	-40

Подходит ли ядро для данных?

Рассмотрим простой искусственный пример.

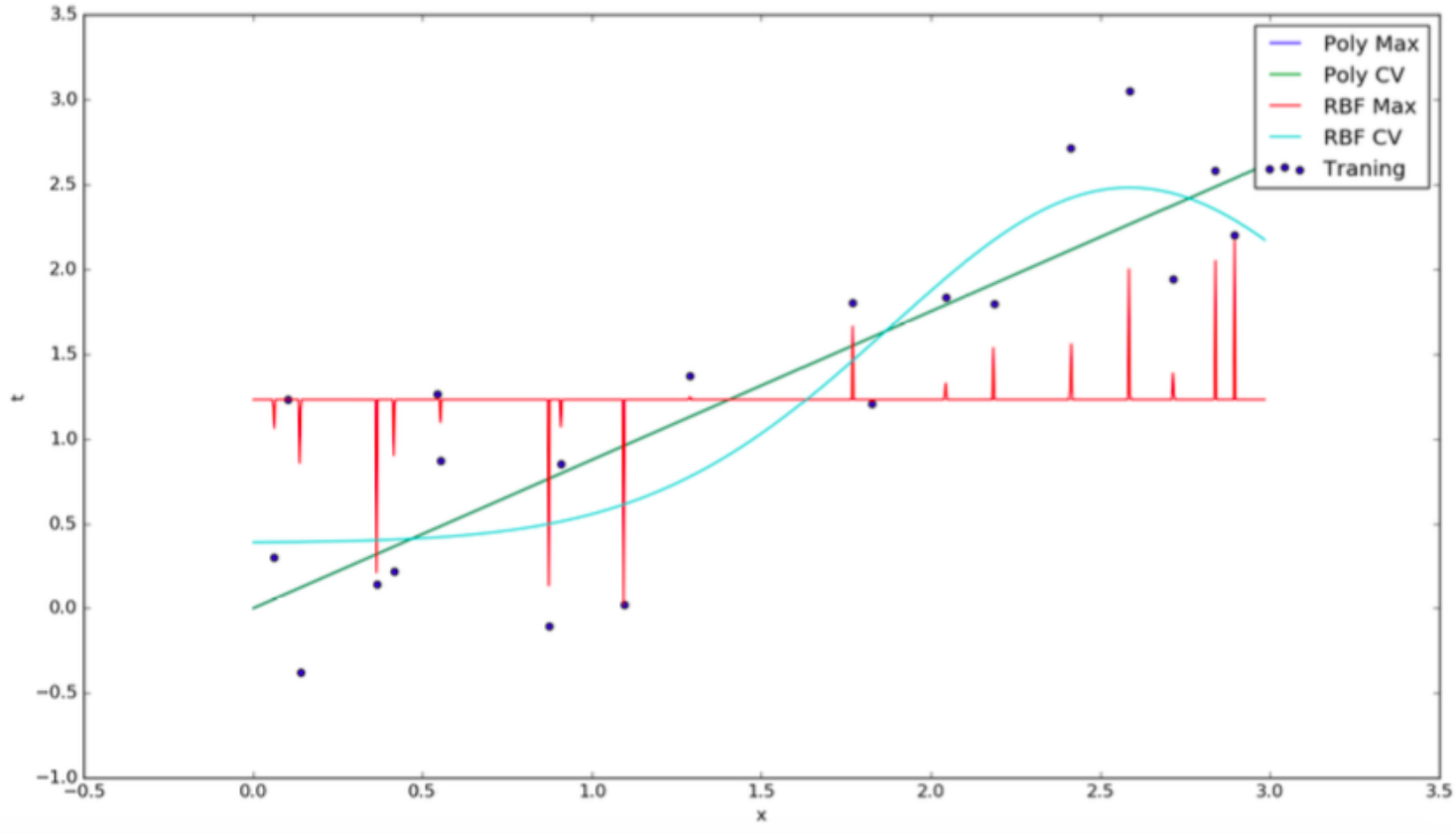
$$\mathrm{t} = \mathrm{x} + \mathcal{N}(0, \sigma^2)$$

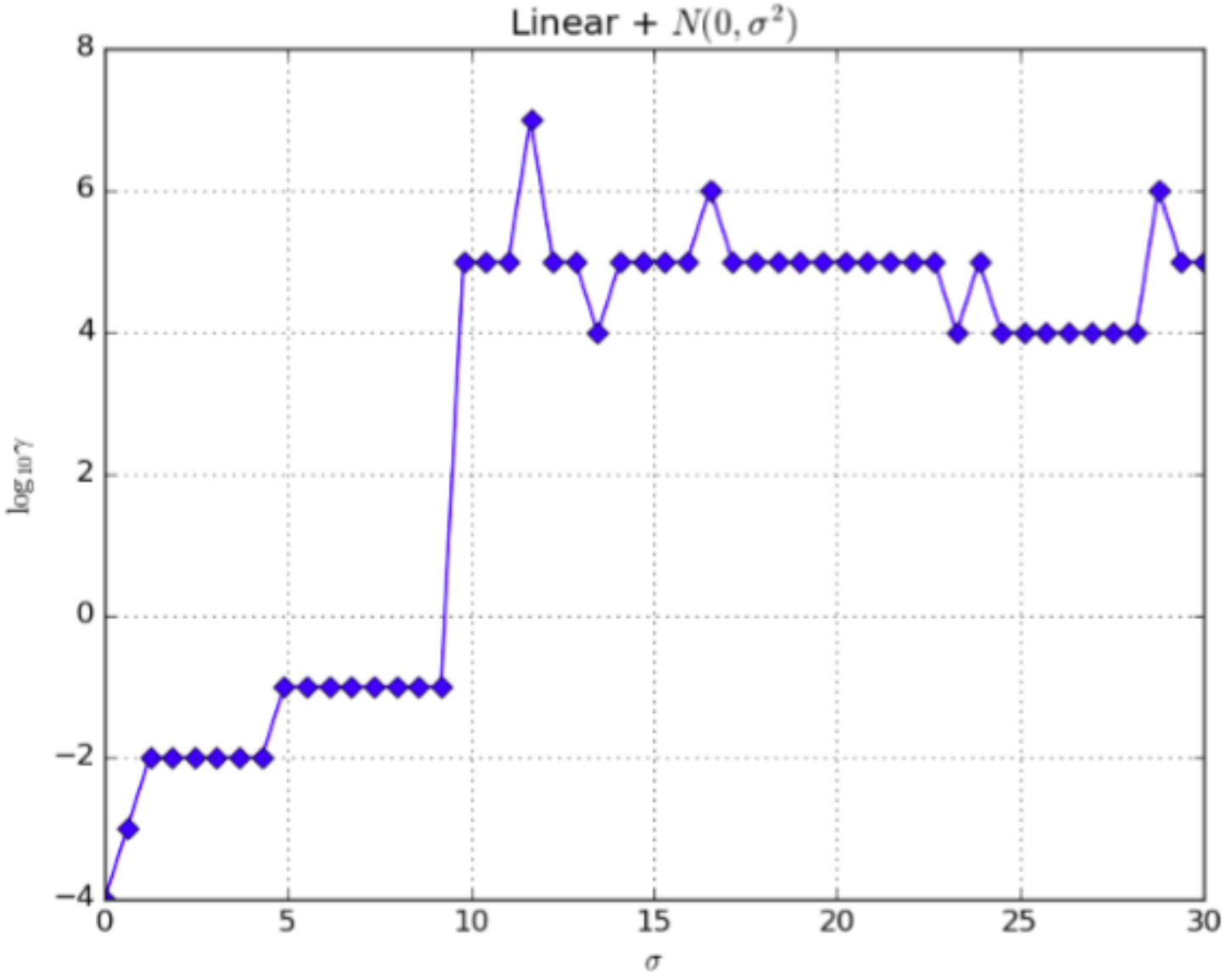


Подходит ли ядро для данных?

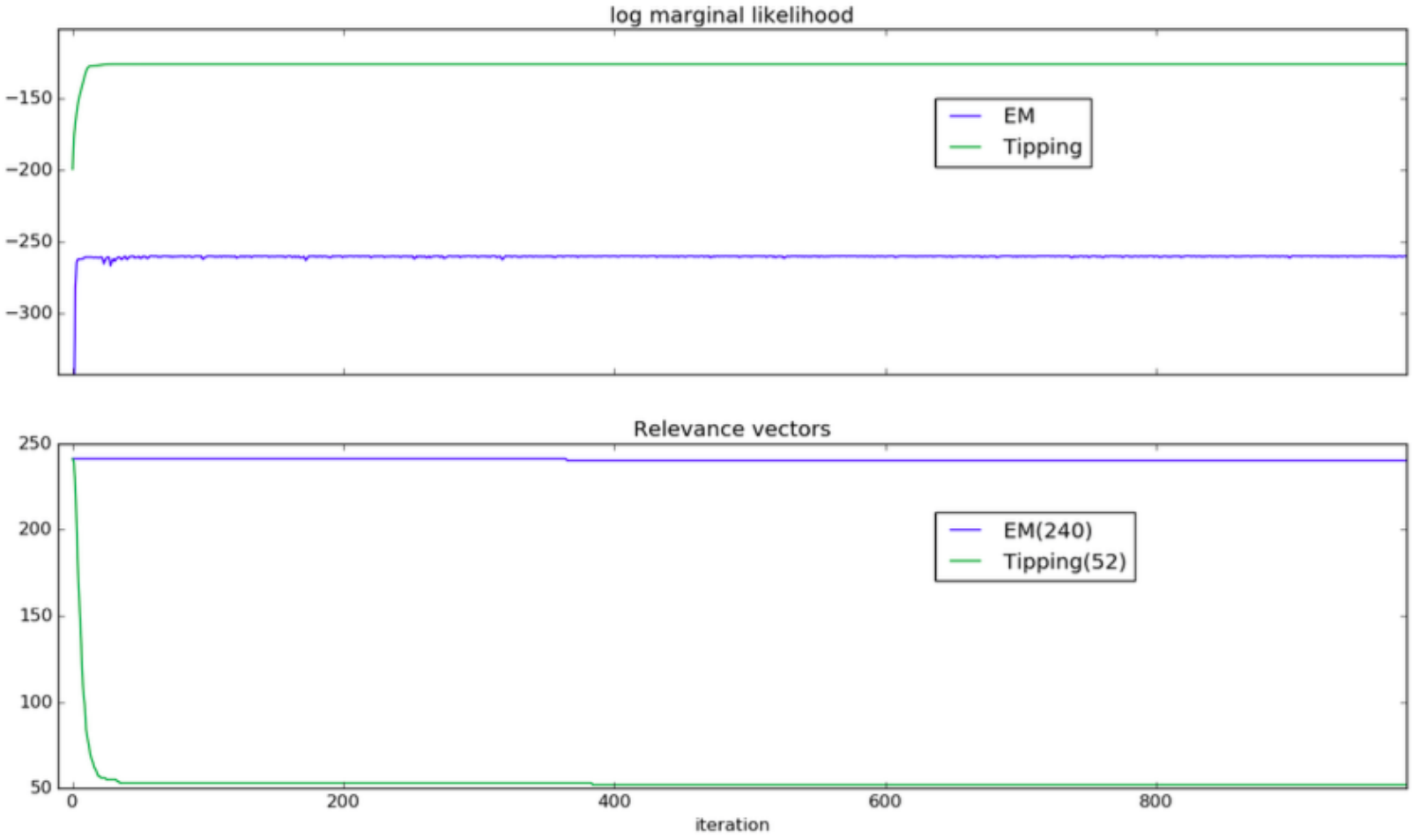
Рассмотрим простой искусственный пример.

$$\mathrm{t} = \mathrm{x} + \mathcal{N}(0, \sigma^2)$$

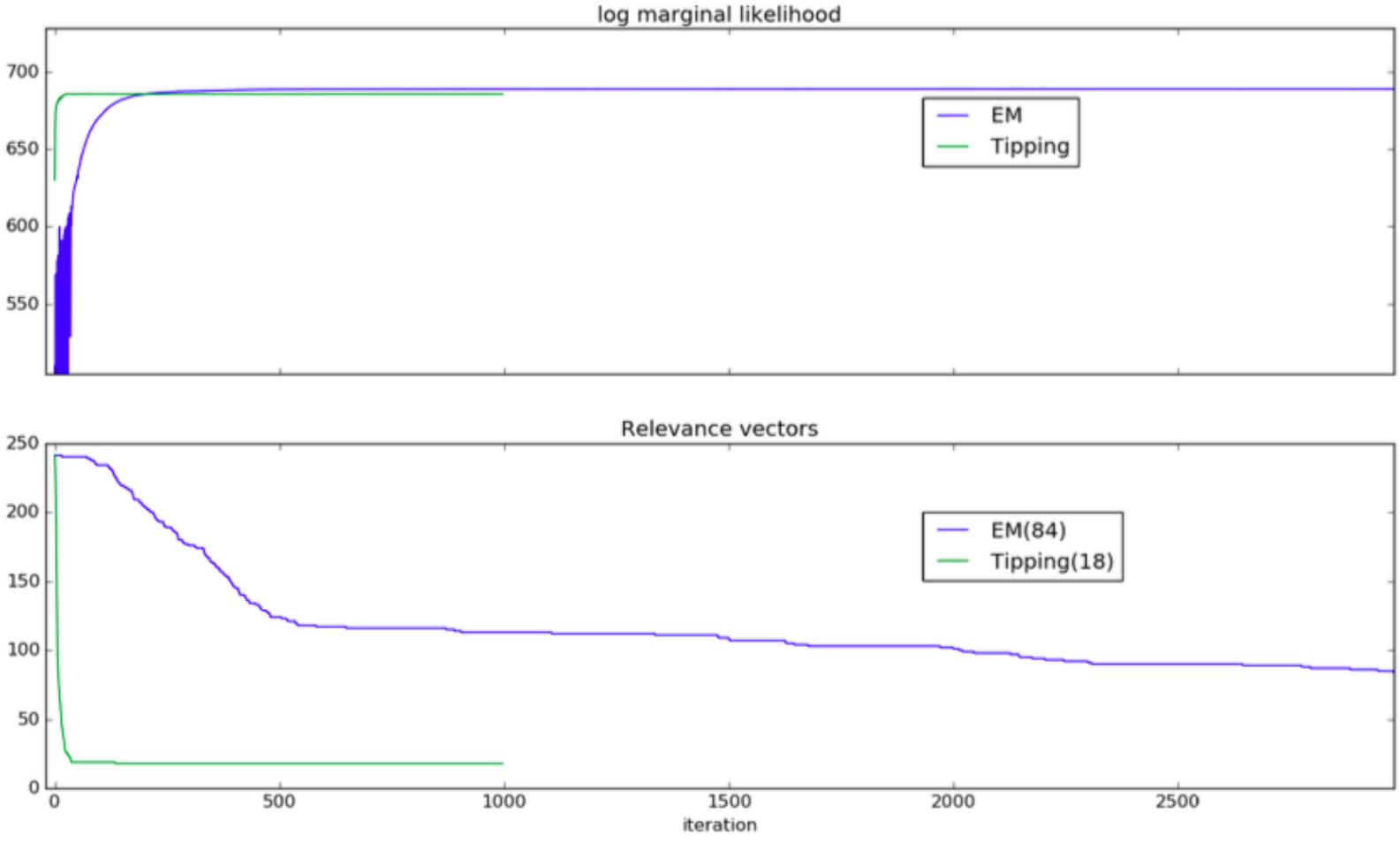



Linear + $N(0, \sigma^2)$
8
6
4
2
0
−2
−4
$\log_{10}\gamma$
0
5
10
15
20
25
30
σ

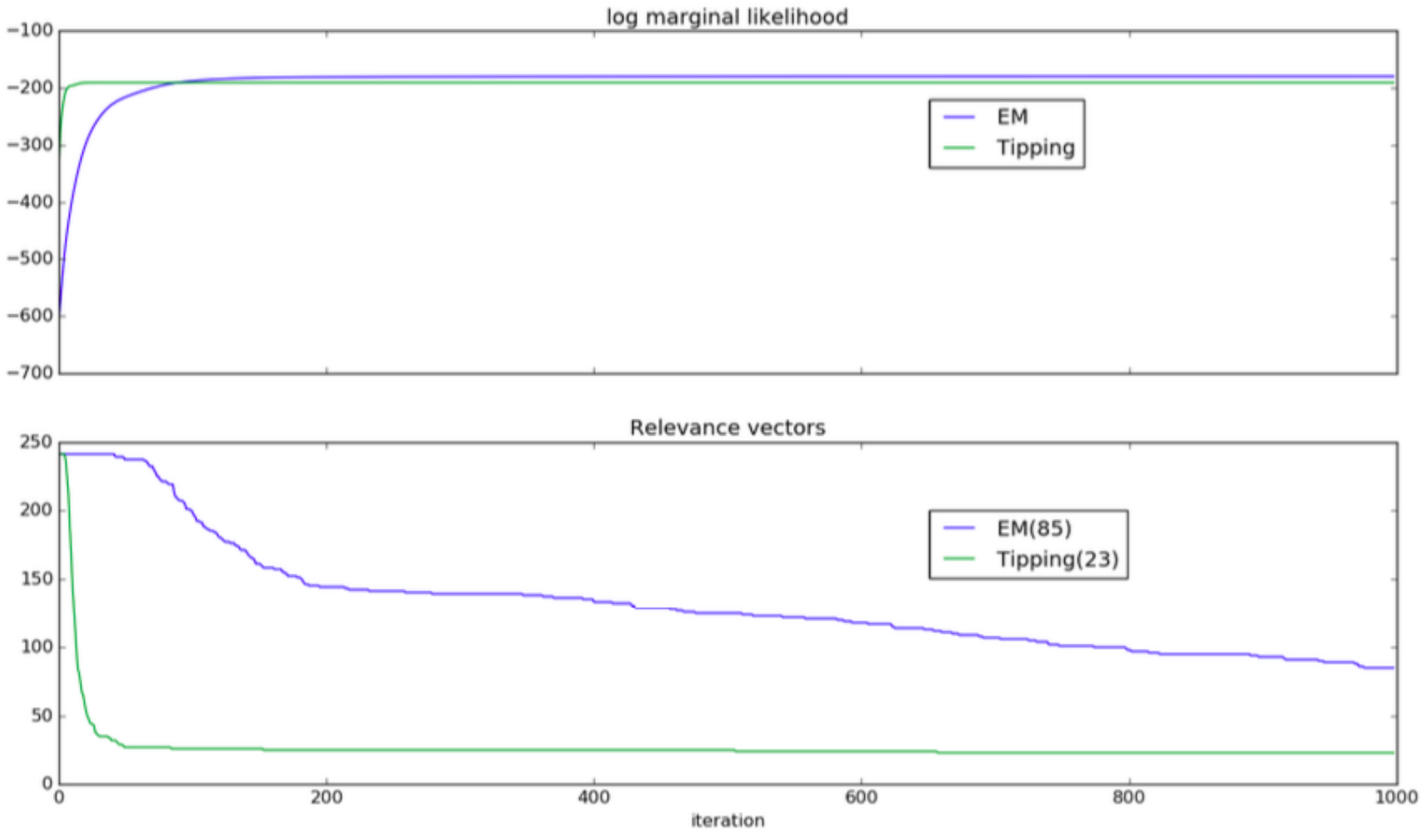
Сходимость


log marginal likelihood
−150
−200
−250
−300
EM
Tipping
Relevance vectors
250
200
150
100
50
0
200
400
600
800
iteration
EM(240)
Tipping(52)

Сходимость


log marginal likelihood
700
650
600
550
EM
Tipping
Relevance vectors
250
200
150
100
50
0
0
500
1000
1500
2000
2500
iteration
EM(84)
Tipping(18)

Сходимость


log marginal likelihood
−100
−200
−300
−400
−500
−600
−700
EM
Tipping
Relevance vectors
250
200
150
100
50
0
EM(85)
Tipping(23)
0
200
400
600
800
1000
iteration

Литература

- Tipping, http://www.jmlr.org/papers/volume1/tipping01a/tipping01a.pdf.
- Bishop - 1.2-1.3, 3.3-3.4, 7.2.
- Лекции - https://www.youtube.com/watch?v=sZxE-BrSMAE&list=PLlb7e2G7aSpR8mbaShVBods-hGaFGifkl

Спасибо за внимание!