Федеральное государственное автономное образовательное
учреждение высшего образования
«Национальный исследовательский университет «Высшая школа экономики»

**Факультет компьютерных наук**
**Основная образовательная программа**
**Прикладная математика и информатика**

# КУРСОВАЯ РАБОТА

на тему

**«Исследование способов подбора базисных функций в методе релевантных векторов»**

Выполнил студент 141 группы, 2 курса,
Атанов Андрей Игоревич

Научный руководитель:
Доцент, Ветров Дмитрий Петрович

**Москва 2016**

# Содержание

**Аннотация**

В работе приводится разбор Метода Релевантных Векторов (Relevance Vector Machine, RVM [1] ). RVM является Байесовским методом машинного обучения, в нем предлагается теоретически обоснованный метод для выбора модели. В работе изложена вероятностная модель лежащая в основе рассматриваемого метода. Основным достоинством метода является часть, отвечающая за автоматический выбор модели - Automatic Relevance Determination (ARD). Основным преимуществом данного метода является отыскание так называемого разреженного решения, которое оставляет лишь небольшое количество «релевантных» признаков в процессе обучения. Помимо теории, в работе также приводятся результаты тестирования алгоритма. Метода максимальной обоснованности, который используется для выбора модели, также можно применить для поиска оптимальных значений параметра в ядровой функции (например, RBF kernel). В работе приводится сравнение данного метода с кросс-валидацией.

**Ключевые слова:** Relevant Vector Machine, Automatic Relevance Determination, Support Vector Machine, Kernel Function

# 1 Введение

RVM для задачи регрессии является линейной моделью, т.е. решающая функция является взвешенной суммой признаков объекта. Функционально модель очень похожа на популярный Support Vector Machine (SVM), но имеет ряд преимуществ. Большой проблемой в машинном обучении остается проблема переобучения. В рамках SVM вводится регуляризация (параметр $C$), которая штрафует большие значения весов. Данная техника помогает бороться с проблемой переобучения, а также отсеивает некоторые признаки. Однако, минусом данного подхода является то, что все веса штрафуются с одним коэффициентом, что в свою очередь сказывается на разреженности найденного решения. RVM предлагает ввести априорное распределение на веса на каждый вес отдельно (что в свою очередь является аналогом регуляризации в SVM, только с разными параметрами). Это приводит к более тщательному отбору релевантных признаков и более разреженному решению.

# 2 Relevance Vector Machine

В этой главе рассматривается математическая модель, лежащая в основе Метода Релевантных Векторов, для задач регрессии и классификации. Все выкладки вынесены в приложение.

## 2.1 Регрессия

### 2.1.1 Постановка задачи

В задаче регрессии на входе имеется выборка $X = \{x_1, ..., x_N\}, x_i \in \mathbb{R}^d$ и вектор объясняемых переменных $t \in \mathbb{R}^n$. Как уже было сказано, регрессионная модель является линейной, таким образом решающая функция имеет вид:

$$y(w, x) = w^T \phi(x)$$

где $\phi(x) = (\phi_1(x), ..., \phi_M(x))^T$ это вектор признаков объекта $x$ из выборки, а $w$ вектор весов. Обычно признаки определяют как ядровые функции, т. е. $\phi_n(x) = K(x, x_n)$, а также добавляется постоянный параметр, который отвечает за смещение (bias), таким образом $M = N + 1$. Стоит также отметить, что на ядровые функции в данной модели накладываются меньшие ограничения нежели в SVM [2]. Полагаем, что:

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

Предполагая, что величины $t_n$ независимы, можно записать функцию правдоподобия выборки $X$:

$$p\left(t|X, w, \sigma^2\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}||t_i - y(w, x_i)||^2\right\} \tag{2}$$

Обозначив $\Phi = (\phi(x_1), ..., \phi(x_N))^T$ – матрица (*design matrix*), где в $i$-ой строке стоят признаки $i$-го объекта, можно переписать (2) в следующем виде:

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

Максимизация напрямую данной величины приведет к неизбежному переобучению, поэтому необходимо ввести регуляризацию, т. е. априорное знание на веса $w$, причем отличительной особенностью данной модели является то, что для веса каждого признака вводится свой параметр регуляризации. Априорное распределение на $w$ выглядит следующим образом:

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

где $A = diag(\alpha_1, ..., \alpha_M)$

### 2.1.2 Выбор параметров

При использовании одного параметра регуляризации, его можно подобрать кроссвалидацией, в данной же модели с учетом того, что обычно $M = N + 1$, данный метод на практике не применим. Однако в рамках Байесовской теории предлагается следующий теоретически обоснованный способ выбора параметров модели.

Оптимальные параметры $\alpha_{MP}, \sigma^2_{MP}$ находятся с помощью метода максимальной обоснованности:

$$\begin{aligned} p\left(t|\alpha,\sigma^2\right) &= \int p(t|\alpha,\sigma^2,w)p(w|\alpha)dw \\ &= \frac{1}{(2\pi)^{-N/2}|\beta I + \Phi A^{-1}\Phi^T|^{1/2}} \exp\left\{-\frac{1}{2}t^T(\beta I + \Phi A^{-1}\Phi^T)^{-1}t\right\} \end{aligned} \tag{3}$$

где $\beta = \sigma^{-2}$

Задача максимизации значения (3) носит название также *type-II maximum likelihood*. Для удобства можно взять логарифм данного значения и записать аналогичную задачу максимизации:

$$\mathcal{L} = -\frac{1}{2}\left(log(|\beta I + \Phi A^{-1}\Phi^T|) + t^T(\beta I + \Phi A^{-1}\Phi^T)^{-1}t\right) \to \max_{\alpha,\beta} \tag{4}$$

Приравнивая градиент целевой в задаче (4) по $\beta$ и $\alpha$ к нулю, получаем следующие формулы пересчета [1] :

$$\alpha_i = \frac{\gamma_i}{w_{MP,i}} \tag{5}$$

$$\beta = \frac{||\Phi w_{MP} - t||^2}{N - \sum_{i=1}^{M}\gamma_i} \tag{6}$$

$$\gamma_i = 1 - \alpha_i \Sigma_{ii}$$

где $\Sigma = (A + \beta\Phi^T\Phi), w_{MP} = \beta\Sigma\Phi^T t$. Подробнее вывод данных формул приводится в приложении.

### 2.1.3 Принятие решения

Для нового объекта вероятность целевой переменной может быть вычислена с помощью формулы полной вероятности по весам $w$:

$$p(t^*|t,\alpha_{MP},\sigma^2_{MP}) = \int p(t^*|w,\alpha_{MP},\sigma^2_{MP})p(w|t,\alpha_{MP},\sigma^2_{MP})dw = \mathcal{N}(t^*|y_*,\sigma^2_*)$$

$$y_* = w^T_{MP}\phi(x)$$
$$\sigma^2_* = \sigma^2_{MP} + \phi(x)^T\Sigma\phi(x)$$

## 2.2 Классификация

Рассмотрим задачу классификации для двух классов, где, как и в регрессии, есть выборка $X$ из $N$ объектов, каждый из которых лежит в $\mathbb{R}^d$, а целевая переменная $t_n$ принимает значения 0 или 1.

Для принятия решения также применяется линейная модель:

$$t_n = sign\left[w^T\phi(x)\right]$$

В качестве вероятности воспользуемся логиситической функцией $\sigma(y) = \frac{1}{1+epx(-y)}$, тогда правдоподобие выборки можно записать следующим образом:

$$P(t|X,w) = \prod_{i=1}^{N} \sigma(w^T\phi(x_n))^{t_i}[1-\sigma(w^T\phi(x_n))]^{1-t_i}$$

Также как и в задаче регрессии введем регуляризацию на весовые коэффициенты $p(w|\alpha) = \mathcal{N}(w|0, A^{-1})$.

Для выбора оптимальных параметров также воспользуемся принципом максимальной обоснованности:

$$p(t|\alpha) = \int p(t|w)p(w|\alpha)dw \to \max_{\alpha}$$

Однако данный интеграл не берется аналитически, как в задаче регрессии, поэтому вычислим его приближенно. Воспользуемся для этого методом Лапласа:

$$\int \exp\{f(x)\}dx \approx (2\pi)^{d/2}|-\nabla\nabla f(x_0)|^{-1/2}\exp\{f(x_0)\}$$

где $x_0$ - стационарная точка, т. е. $\nabla f(x_0) = 0$.

Для фиксированных $\alpha$ вектор «наиболее вероятных» весов также необходимо искать в приближенной форме. Для этого можно воспользоваться методом второго порядка Ньютона для следующей задачи оптимизации:

$$\ln\{P(t|w)p(w|\alpha)\} = \sum_{i=1}^{N}\left[t_i ln\left(\sigma_i\right) + (1-t_i) ln\left(1-\sigma_i\right)\right] - \frac{1}{2}w^T Aw \to \max_{w}$$

где $\sigma_i = \sigma(w^T\phi(x_i))$.

Для нахождения максимума обоснованности получаем такие же формулы пересчета (5) и (6), что и для регрессии, за тем лишь исключением, что:

$$\Sigma = \left(\Phi^T B\Phi + A\right)^{-1}$$

$$B = diag(\beta_1, ..., \beta_N), \beta_i = \sigma_i(1-\sigma_i).$$

## 2.3 Вывод

В процессе обучения большинство $\alpha$ будет стремиться к $+\infty$ что приводит к тому, что соответствующие веса обнуляются. Таким образом, за счет индивидуальных коэффициентов регуляризации большое количество векторов отсеиваются в процессе переобучения и решение получается намного разреженнее, нежели в SVM [1]. Однако такой подход может плохо сказаться на обобщающей способности алгоритма, привести к переобучению в простых ситуациях [3]. Основным минусом RVM по сравнению с SVM является вычислительная сложность алгоритма обучения. Однако стоит также учитывать, что для SVM необходимо подбирать параметр регуляризации $C$ с помощью кросс-валидации. Кроме того, ввиду разреженности решения RVM, процесс классификации (восстановления регрессии) требует меньшего количество времени.

# 3 Выбор параметров

Несмотря на то, что в рассматриваемой модели параметры $\alpha$ и $\beta$ подбираются автоматически, необходимо задать их начальное приближение. Помимо этого во многих ядровых функциях есть параметры, которые выставляются эвристически. В этой главе сравниваются способы выбора параметров, а также некоторые оптимальные их значения.

## 3.1 Начальные приблежения

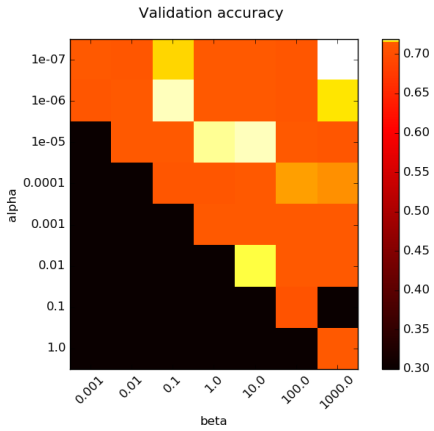



Рис. 1: Зависимость качества[1] построенной регрессии на кросс-валидации для начальных приближений $\alpha$[2] и $\beta$.

[1] В качестве оценки использовался коэффициент детерминации $R^2 = 1 - SS_{res}/SS_{tot}$

[2] Начальное приближение ищется в виде $\alpha\mathbf{1}$

В первой главе был предложен способ автоматического выбора параметров, который также носит название ARD (Automatic Relevance Determination). Но возникает вопрос с какого начального приближения запускать пересчет параметров. На Рис. 1 приведен пример почему это может быть важно. На графике хорошо видно, что качество регрессии сильно разнится. В худшем случае необходимо подбирать приближения для каждого набора данных, что приводит к росту сложности алгоритма.

В Таблице 1 приведены результаты кросс-валидации для выбора оптимальных значений начальных приближений с целью установить, можно ли найти универсальные значения для разных данных на примере 4 наборов.

| Данные | $\alpha$ | $\beta$ |
|---|---|---|
| Friedman # 1 | $10^{-7}$ | $10^{4}$ |
| Friedman # 2 | $10^{-7}$ | 1 |
| Friedman # 3 | $10^{-5}$ | 0.01 |
| Boston Housing | $10^{-8}$ | 0.1 |

Таблица 1: Оптимальные начальные приближения.

В целом видно, что значение $\alpha \approx 10^{-7}$ является оптимальным. В то же время для $\beta$ разброс оптимальных значений достаточно велик, и, как следствие, требует подбора.

Напомню, что до текущего момента никаких манипуляций с данными не проводилось. Попробуем пронормировать выборку, сделав значение выборочной дисперсии равным единице (т. е. поделим на стандартное отклонение), и посмотреть на оптимальный выбор $\beta$. Далее в качестве $\alpha$ используется значение $10^{-7}$.

| Данные | $\gamma$ | $\beta$ |
|---|---|---|
| Friedman #1 | 0.01 | 0.01 |
| Friedman #2 | 0.01 | 0.001 |
| Friedman #3 | 0.1 | 0.001 |
| Boston Housing | 0.001 | 0.001 |

Таблица 2: Оптимальные значения параметров для нескольких задачь регрессии

В Таблице 2 находится также оптимальное значение параметра $\gamma$ для ядровой функции RBF ($K(x, y) = \exp\{-\gamma||x - y||^2\}$). По этим показателям уже можно делать вывод о наличии универсального значения $\beta$ для различных наборов данных. При этом, также стоит отметить, что значение качества регрессии на кроссвалидации не сильно отличается (Табл. 3), и, как следствие, можно брать, вообще говоря, любое приближение.

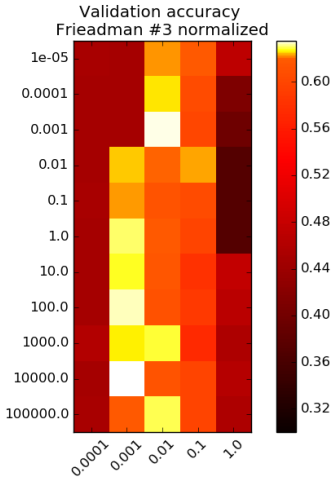



Рис. 2: Пример настройки параметров $\beta$(по вертикали), $\gamma$(по горизонтали) для *Friedman#3*

| $\beta$ | 0.0001 | 0.001 | 0.01 | 0.1 | 1 | 10 | 100 |
|---|---|---|---|---|---|---|---|
| $R^2$ | 0.853 | 0.863 | 0.877 | 0.865 | 0.874 | 0.867 | 0.866 |

Таблица 3: Показатель коэффициента детерминации для разых начальных приближений $\beta$

## 3.2 Параметры ядровой функции

В первой главе было показано, как с помощью метода максимальной обоснованности выбирать $\alpha$. Также упоминалась необходимость выбора параметров ядровой функции с помощью кросс-валидации, которая в свою очередь является ресурсоемкой операцией по времени выполнения. В данном разделе сравнивается качество выбора параметров ядровой функции с помощью метода максимальной обоснованности и кросс-валидации.

Как видно из Табл. 4, в ряде тестов максимизация обоснованности оказывается не хуже, а иногда даже превосходит по некоторым показателям кросс-валидацию. Для полиномиальной ядровой функции на всех датасетах показатели практически не отличаются, однако для RBF на двух последних наборах данных наблюдается резкое ухудшение качества регрессии. На Friedman #3 параметр $\gamma$ устремляется к левой границе сетки перебор ($10^4$), что соответствует выбору крайне узких гауссиан, и, как следствие, модель переобучивается. Этого не происходит с полиномиальным ядром ввиду его простоты.

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

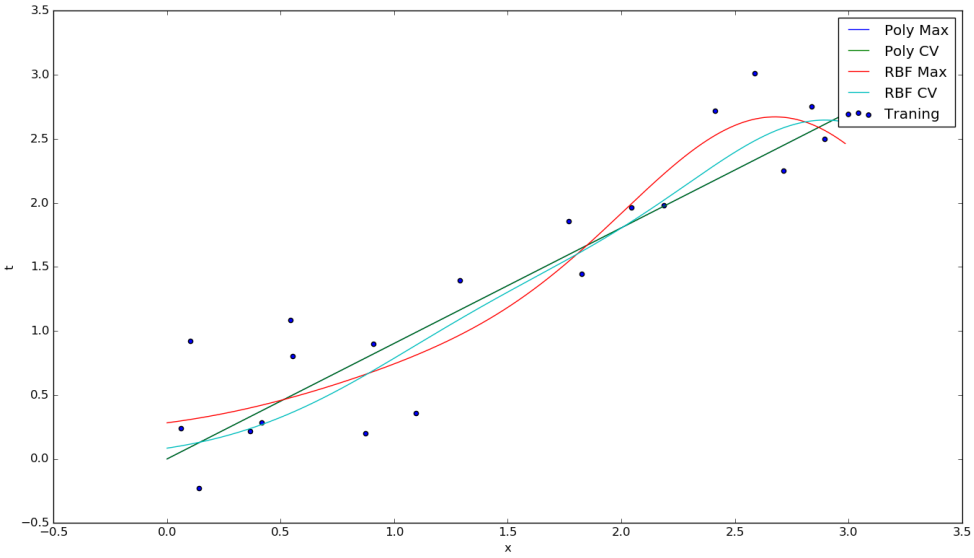
| Data set | Polynomial Kernel | | | RBF Kernel | | |
|---|---|---|---|---|---|---|
| | CV $R^2$ | Test $R^2$ | $\log p(t\|\alpha, \sigma^2)$ | CV $R^2$ | Test $R^2$ | $\log p(t\|\alpha, \sigma^2)$ |
| Airfoil | 0.495 | 0.591 | -221 | 0.666 | 0.739 | -168 |
| Concrete | 0.656 | 0.722 | -152 | 0.663 | 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 |

Таблица 4: В таблицах представлены такие параметры регрессии, как коэффициент детерминации на кросс-валидации (CV $R^2$), на тестовой выборке (Test $R^2$), логарифм обоснованности ($\log p(t|\alpha, \sigma^2)$). Регрессии строились с использованием двух ядровых функций, в первой таблице параметры подбирались с помощью кросс-валидации, во второй с помощью максимизации обоснованности

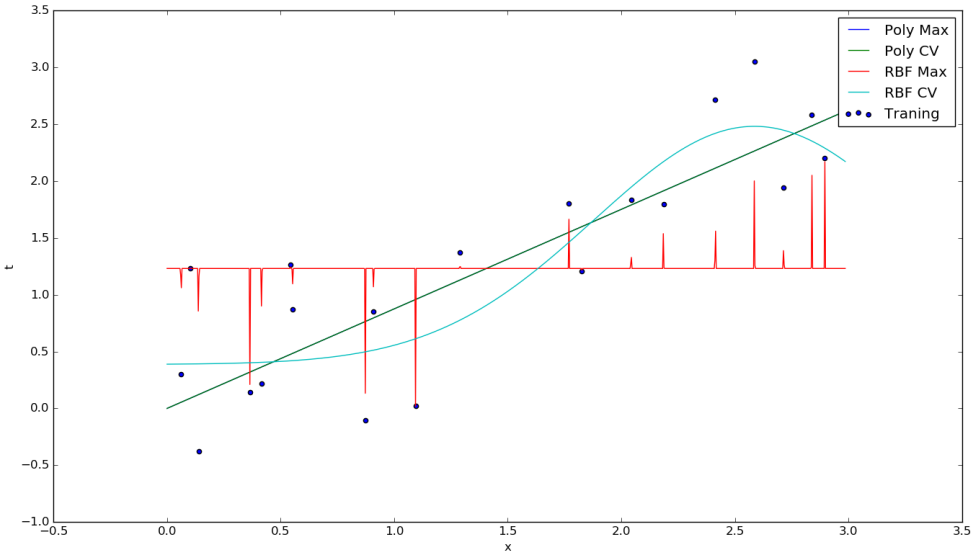
Важно отметить, что при тестировании модели на Friedman #3 к целевым переменным добавлялся нормальный шум $\epsilon \sim \mathcal{N}(0, \sigma^2), \sigma = \frac{1}{8}$. Если же не добавлять шум, оба метода дадут похожие результаты. Принимая данный факт во внимание, этот эффект скорее стоит называть не переобучением, а отказ от нахождения взаимосвязи в данных. Таким образом алгоритм говорит о том, что нет зависимости в данных. Рассмотрим данный эффект подробнее в следующем разделе.

## 3.3 Зависимость в данных

Рассмотрим более подробно эффект, отмеченный в предыдущем разделе, на следующем синтетическом примере: $t_n = x_n + \sigma\mathcal{N}(0, 1)$.




(a) $\sigma = 20$




(b) $\sigma = 30$

Рис. 3: На графике отображено множество точек из обучающей и тестовой выборках, а также построенные регрессии с соответствующими ядрами. Параметры функций определялись двумя способами.

На Рис. 3 представлено два графика, различающиеся дисперсией шума, на первом графике наиболее обоснованный параметр оказывается близок к тому, что

выбирается кросс-валидацией. Увеличивая дисперсию, получаем ситуацию, проиллюстрированную на втором графике. Здесь наиболее обоснованный параметр соответствует сверхузким гауссианам и регрессионная модель просто запоминает выборку.

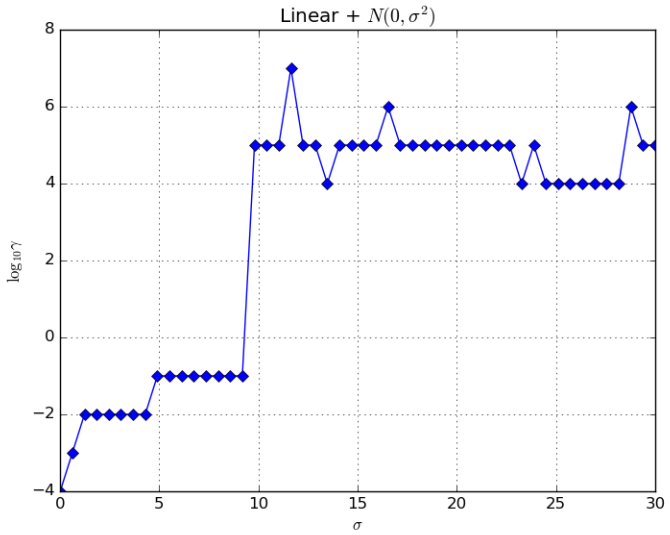



Рис. 4: График отражает зависимость логарифма параметра $\gamma$ ядровой функции RBF от стандартного отклонение нормального шума, добавляемого к данным.

Видно, что на Рис. 3b при использовании ядровой функции RBF регрессионная модель описывает данные совсем не верно, хотя и ошибка на кросс-валидации, и ошибка на тесте у нее будет меньше.

Таким образом, выбор сверх узких гауссиан может быть гораздо полезнее и информативнее, нежели попытка выбора гауссиан, хоть сколько-нибудь объясняющих данные. Такой результат свидетельствует о том, что данная ядровая функция не подходит под описание данных и стоит выбрать другую.

## 3.4 Вывод

В этой главе были представлены количественные эксперименты, с помощью которых были установлены оптимальные значения начального приближения $\alpha$ 1 и $\beta$ в случе нормированной выборки 2. Помимо этого, был предложен способ выбора параметра ядровой функции с помощью метода максимальной обоснованности.

# 4 Сравнение формул пересчета

В первой главе приводились формулы пересчета параметров $\alpha$ и $\beta$ – (5) и (6). Применяя EM алгоритм [1], можно получить похожие формулы:

$$\alpha = \frac{1}{\gamma_i^2 + \Sigma_{ii}} \tag{7}$$

$$\beta^{new} = \frac{N}{||t - \Phi\mu||^2 + (\beta^{old})^{-1} \sum_{i=1}^{M} \gamma_i} \tag{8}$$

В этой главе приводятся результаты сравнения двух вариантов формул на некоторых наборах данных. В ходе экспериментов сравнивались такие показатели, как количество найденных релевантных векторов, значение обоснованности и один из самых главных показателей – скорость схождения. Поскольку на каждой итерации приходится обращать матрицу размера $M \times M$, очень важно добиться быстрой сходимости метода, иначе на больших выборках алгоритм будет работать непомерно долго.

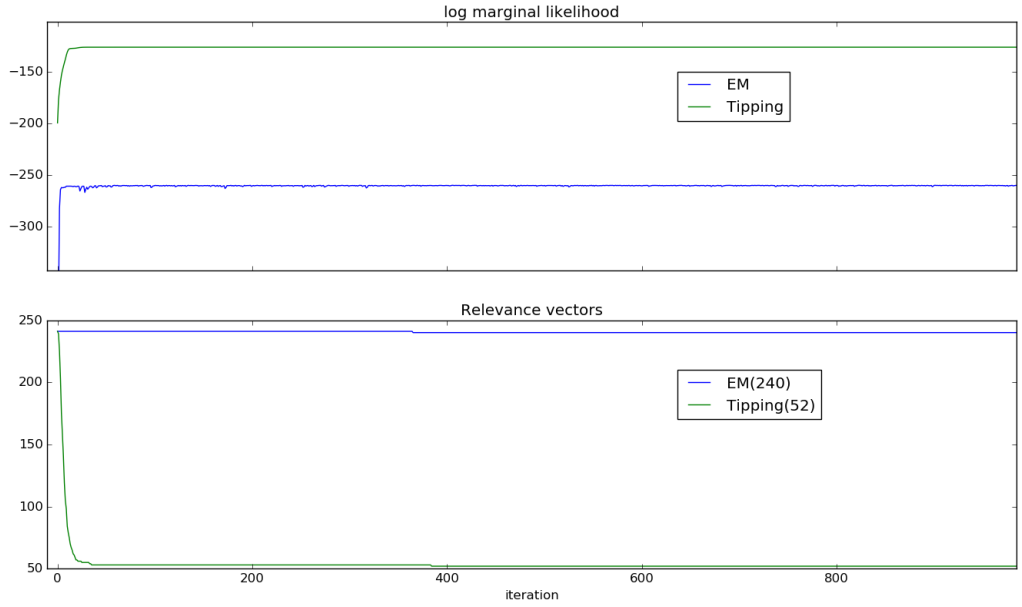



Рис. 5: График отражает значение логарифма обоснованности и количество релевантных векторов на каждой итерации пересчета для набора данных Friedman #1

На Рис. 5 оба метода сходятся достаточно быстро, однако видно, что EM находит гораздо меньшее значение обоснованности. При этом качество обеих регрессий достаточно хорошее – $R^2 \approx 0.9$. На нижнем графике видно, что EM оставил все

вектора релевантными, так что ни о какой разряженности решения говорить в данном случае не приходится.

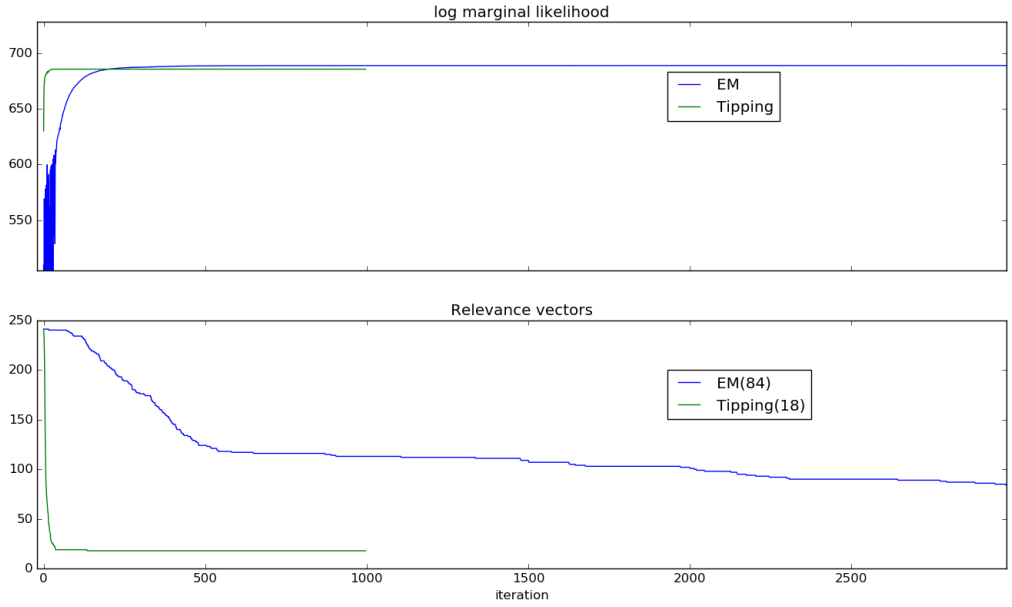



Рис. 6: График отражает значение логарифма обоснованности и количество релевантных векторов на каждой итерации пересчета для набора данных Friedman #2

На Friedman #2 Рис. 6 EM ведет себя лучше чем в предыдущем примере, однако ему требуется гораздо большее количество итераций, причем количество релевантных векторов в несколько раз больше, чем для формул, предложенных Типпигом [1]. На Рис. 7 видно, что формулы Типпинга сходится уже после 50 итераций. Стоит также отметить флуктуации в показаниях обоснованности на первых шагах EM.

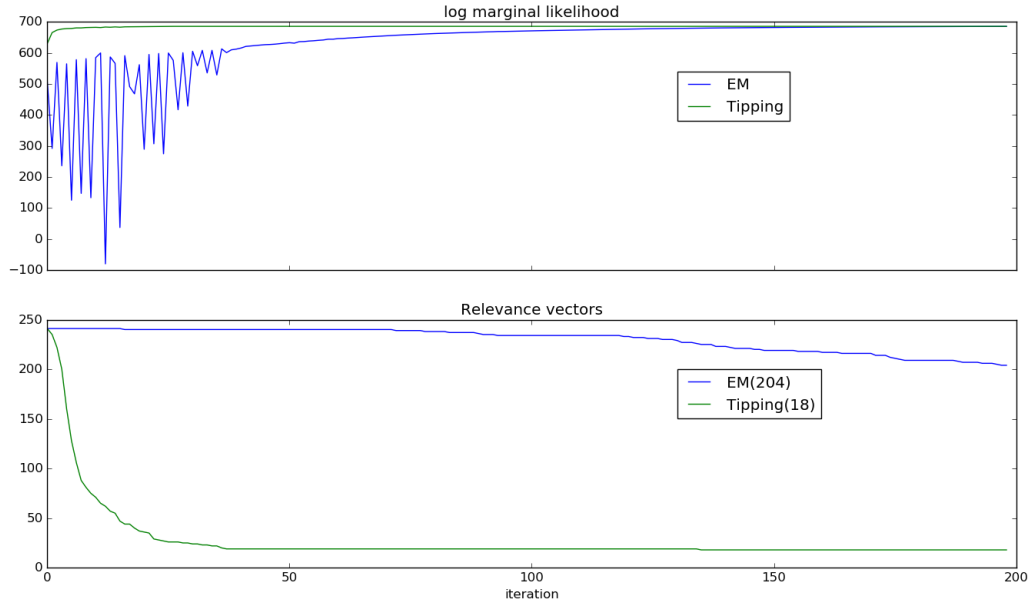



Рис. 7: График отражает значение логарифма обоснованности и количество релевантных векторов на каждой итерации пересчета для набора данных Friedman #2

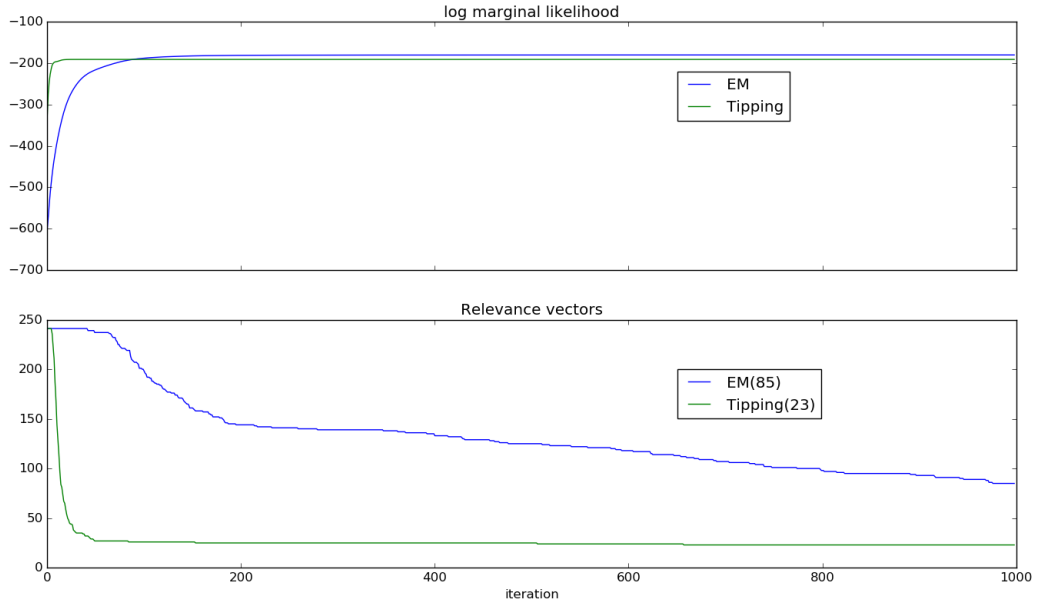



Рис. 8: График отражает значение логарифма обоснованности и количество релевантных векторов на каждой итерации пересчета для набора данных Friedman #3

Помимо быстрой сходимости и нахождения разряженного решения, не менее важным плюсом формул Типпинга является то, что они сходятся монотонно на всех приведенных примерах. Если для EM монотонная сходимость является фактом теоретическим (хотя на Рис. 7 видны сильные флуктуации), то для Типпинга ничего такого не гарантируется, и данное наблюдение является крайне полезным.

Таким образом формулы пересчета параметров Типпинга не только быстро сходятся, но и обеспечивают монотонный рост, а также находят необходимое решение, тем самым давая возможность применять данный метод обучения на более больших объемах данных.

# 5 Заключение

В данной работе мною был изучен метод машинного обучения Relevance Vector Machine. Помимо изучения теории, результатом работы также является программная реализация алгоритма регрессии и классификации.

**Основными достоинствами RVM являются:**

1. Возможность выбора любых признаков, в отличие от SVM, в котором ядровые функции должны удовлетворять условию положительной определенности.

2. Автоматический выбор параметров модели с помощью Байесовского подхода.

3. Возможность выбора параметра ядровой функции с помощью максимизации обоснованности. Кроме того, как обсуждалось ранее 3.3, такой подход дает не только выигрыш по времени (требуется лишь один раз обучиться, чтобы вынести оценку в виде значения обоснованности), но и умеет определять, есть ли в данных зависимость, соответствующая выбранному ядру.

4. Разреженность решения, за счет подбора для каждого веса своего параметра регуляризации.

К минусам можно отнести сложность алгоритма (на каждой итерации пересчета параметров $\alpha$ и $\beta$ требуется обращать матрицу размера $M \times M$), а также приближенный характер решения.

В дальнейшем необходимо будет провести соответствующие эксперименты для задачи классификации. Основной задачей в дальнейшем является реализация стохастической оптимизации для нахождения максимума обоснованности, а также масштабирование метода на случай выборок больших объемов.

# Приложение А Пересчет параметров

В первой главе был предложен способ выбора параметров с помощью максимизации обоснованности. В данном приложении будет более подробно рассмотрен вывод формул пересчета.

Имеется следующая задача оптимизации:

$$\mathcal{L} = -\frac{1}{2}\left(\log(|\beta I + \Phi A^{-1}\Phi^T|) + t^T(\beta I + \Phi A^{-1}\Phi^T)^{-1}t\right) \to \max_{\log\alpha,\log\beta} \tag{9}$$

Максимизация по $\alpha, \beta$ и $\log\alpha, \log\beta$ эквивалентны. Прежде чем находить градиент, преобразуем целевую функцию.

Воспользуемся матричным тождеством Шермана-Моррисона-Вудбери и преобразуем второе слагаемое в (9)

$$(\beta I + \Phi A^{-1}\Phi^T)^{-1} = \beta I - \beta\Phi(A + \beta\Phi^T\Phi)\Phi^T\beta$$

$$t^T(\beta I + \Phi A^{-1}\Phi^T)^{-1}t = \beta t^T(t - \Phi\mu)$$

где $\mu = \beta\Sigma\Phi^T t$ – апостериорное среднее весовых коэффициентов.

$$\begin{aligned}
\beta t^T(t - \Phi\mu) &= \beta(t^T - \mu^T\Phi^T + \mu^T\Phi^T)(t - \Phi\mu) \\
&= \beta||t - \Phi\mu||^2 + \beta t^T\Phi\Sigma\Sigma^{-1}\mu - \beta\mu^T\Phi\Phi^T\mu \\
&\left[\Sigma^T = \Sigma,\ \beta t^T\Phi\Sigma = \mu^T\right] \\
&= \beta||t - \Phi\mu||^2 + \mu^T\left(\Sigma^{-1} - \beta\Phi\Phi^T\right)\mu \\
&= \beta||t - \Phi\mu||^2 + \mu^T A\mu
\end{aligned} \tag{10}$$

Далее воспользуемся следующим тождеством для определителя матрицы:

$$|A + UV| = |I + VA^{-1}U||A|$$

Преобразуем с его помощью первое слагаемое:

$$\begin{aligned}
\log(|\beta I + \Phi A^{-1}\Phi^T|) &= -N\log\beta + \log|A + \beta\Phi^T\Phi| - \log|A| \\
&= -N\log\beta + \log|\Sigma^{-1}| - \log|A|
\end{aligned} \tag{11}$$

## А.1 Параметр $\alpha_i$

Далее вычислим частную производную по $\log\alpha_i$:

$$\frac{\partial\log|\Sigma^{-1}|}{\partial\log\alpha_i} = \frac{M_{ii}}{\log|\Sigma|}\frac{\partial\Sigma_{ii}^{-1}}{\partial\log\alpha_i} = \Sigma_{ii}\alpha_i$$

$$\frac{\partial\log|A|}{\partial\log\alpha_i} = 1$$

$$\frac{\partial \beta t^T (t - \Phi\mu)}{\log \alpha_i} = -\beta t^T \Phi \frac{\partial \mu}{\partial \log \alpha_i} = \beta t^T \Phi \beta \frac{\partial \Sigma}{\partial \log \alpha_i} \Phi^T t$$

$$\frac{\partial \Sigma}{\partial \log \alpha_i} = -\Sigma \frac{\partial (A + \beta \Phi^T \Phi)}{\partial \log \alpha_i} \Sigma = -\Sigma \frac{\partial A}{\partial \log \alpha_i} \Sigma = \alpha_i \Sigma_{ii}^2$$

Собирая все воедино и приравнивая производную к нулю, получаем:

$$\frac{1}{2} \left(1 - \alpha_i \left(\mu_i^2 + \Sigma_{ii}\right)\right) = 0$$

$$\alpha_i = \frac{\gamma_i}{\mu_i}$$

## A.2 Параметр $\beta$

Найдем теперь производную $\mathcal{L}$ по $\log \beta$. Сделаем это также в несколько этапов.

Воспользуемся следствием из формулы Якоби для производной определителя матрицы и вычислим частную производную $|\Sigma^{-1}|$:

$$\frac{\partial \log |\Sigma^{-1}|}{\partial \log \beta} = \beta \ tr\big(\Sigma \Phi^T \Phi\big)$$

Далее найдем производную первого слагаемого в (10)

$$\frac{\partial \beta ||t - \Phi\mu||^2}{\partial \log \beta} = \beta ||t - \Phi\mu||^2 + \beta \frac{\partial ||t - \Phi\mu||^2}{\partial \log \beta} \tag{12}$$

$$\begin{aligned} \frac{\partial ||t - \Phi\mu||^2}{\partial \log \beta} &= \frac{\partial (t - \Phi\mu)^T (t - \Phi\mu)}{\partial (t - \Phi\mu)} \ \frac{\partial (t - \Phi\mu)}{\partial \log \beta} \\ &= 2 \left(t - \Phi^T \mu\right) \left(-\Phi \frac{\partial \mu}{\partial \log \beta}\right) \end{aligned} \tag{13}$$

$$\frac{\partial \mu}{\partial \log \beta} = \frac{\partial \beta \Sigma \Phi^T t}{\partial \log \beta} = \begin{pmatrix} \partial \frac{1}{\beta} \Sigma^{-1} \\ -\beta \Sigma \frac{\partial \frac{1}{\beta} \Sigma^{-1}}{\partial \log \beta} \beta \Sigma \end{pmatrix} \Phi^T t = \beta \Sigma A \mu$$

Подставляя данное выражение сначала в (13), а затем полученное выражение в (12), получаем:

$$\frac{\partial ||t - \Phi\mu||^2}{\partial \log \beta} = \beta ||t - \Phi\mu||^2 - 2\beta^2 t^T \Phi \Sigma A \mu + 2\beta^2 \mu \Phi^T \Phi \Sigma A \mu \tag{14}$$

Далее найдем производную второго слагаемого в (10), используя ранее полученный результат:

$$\frac{\partial \mu^T A \mu}{\partial \log \beta} = 2\mu^T A \frac{\partial \mu}{\partial \log \beta} = 2\beta \mu^T A \Sigma A \mu \tag{15}$$

Рассмотрим последнее слагаемое в (14) и (15):

$$2\beta^2 \mu \Phi^T \Phi \Sigma A \mu + 2\beta \mu^T A \Sigma A \mu = 2\beta \mu \left(A + \beta \Phi^T \Phi\right) \Sigma A \mu = 2\beta \mu \Sigma^{-1} \Sigma A \mu = 2\beta \mu A \mu$$

Заметим, что это в точности второе слагаемое в (14), взятое с обратным знаком. Таким образом получаем:

$$\frac{\partial \beta ||t - \Phi \mu||^2}{\partial \log \beta} = \beta ||t - \Phi \mu||^2$$

В итоге, приравнивая к нулю найденную производную, получаем следующее уравнение:

$$\frac{N}{\beta} ||t - \Phi \mu||^2 - tr\left(\Sigma \Phi^T \Phi\right)$$

Принимая во внимания равенство $tr\left(\Sigma \Phi^T \Phi\right) = \beta^{-1} \sum_{i=1}^{M} \gamma_i$ [1], получаем следующую формулу пересчета для параметра $\beta$:

$$\beta = \frac{N - \sum_{i=1}^{M} \gamma_i}{||t - \Phi \mu||^2}$$

# Приложение В Метод Лапласа

В первой главе в задаче классификации для вычисления обоснованности использовался метод Лапласа приближения интеграла вида $\int \exp f(x)dx$. В данном приложении приведены более подробные выкладки по применению данной аппроксимации, а также по нахождению вектора оптимальных весов $w_{MP}$.

Для обоснованности имеем:

$$f(w) = \sum_{i=1}^{N} \left[t_i \ln(\sigma_i) + (1-t_i)\ln(1-\sigma_i)\right] - \frac{1}{2}w^T A w$$

Сначала с помощью метода Ньютона второго порядка найдем стационарную точку, для этого найдем градиент и матрицу Гесе данной функции:

$$\nabla f(w) = \sum_{i=1}^{N} (t_i - \sigma_i)\,\phi(x_i) - Aw = \Phi^T(t-y)$$

$\frac{\partial\sigma(x)}{\partial x} = \sigma(1-\sigma), y = (\sigma_1, ..., \sigma_N)$.

$$\nabla\nabla f(w) = -\Phi^T \frac{\partial y}{\partial w} - A = -(\Phi^T B \Phi + A)$$

Таким образом получаем следующую формулу для пересчета вектора весов:

$$w_{new} = w_{old} + (\Phi^T B\Phi + A)^{-1}\Phi^T(t-y)$$

Далее воспользуемся приближением гауссианной:

$$\ln\{p(t|\alpha)\} \approx \sum_{i=1}^{N} \left[t_i \ln(\sigma_i) + (1-t_i)\ln(1-\sigma_i)\right] - \frac{1}{2}\left[w_{MP}^T A w_{MP} + \ln|\Sigma| + \ln|A|\right]$$

где $\Sigma = \left(\Phi^T B\Phi + A\right)^{-1}, B = diag(\beta_1, ..., \beta_N), \beta_i = \sigma_i(1-\sigma_i)$.

Заметим, что сумма не зависит от $\alpha$, тогда:

$$\frac{\partial \ln\{p(t|\alpha)\}}{\partial \ln \alpha_i} = \frac{1}{2}\left[1 - \alpha_i\left(w_{MP}^2 + \Sigma_{ii}\right)\right]$$

Приравнивая производную к нулю получаем такие же формулы пересчета, как и для задачи регрессии.

# Список литературы

[1] Tipping Michael E. Sparse Bayesian Learning and the Relevance Vector Machine // J. Mach. Learn. Res. 2001. Т. 1. С. 211–244.

[2] Bishop Christopher M. Pattern Recognition and Machine Learning (Information Science and Statistics). Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[3] Predictive Automatic Relevance Determination by Expectation Propagation / Yuan (Alan) Qi, Thomas P. Minka, Rosalind W. Picard [и др.] // Proceedings of the Twenty-first International Conference on Machine Learning. ICML '04. New York, NY, USA: ACM, 2004. С. 85–. URL: http://doi.acm.org/10.1145/1015330.1015418.