Гауссовские процессы для активного обучения в задаче классификации

Дарья Котова[1], Максим Панов[2]

1: Московский физико-технический институт (ГУ)
kotova.ds@phystech.edu
2: Сколковский институт науки и технологий
m.panov@skoltech.ru

Аннотация Активное обучение – новый подход к обучению моделей, который основывается на адаптивном увеличении обучающей выборки ранее неразмеченными точками. Как правило выбор точек основан на оценке неопределенности предсказания модели в точки. Гауссовский процесс – стохастический процесс, такой что любой конечный набор этих случайных величин имеет многомерное нормальное распределение.
Недавно был предложен новый критерий для алгоритма активного обучения [1], с приложением так же и для гауссовских процессов. Критерий оценивает норму функции, которая описывает тренировочное множество в совокупности с еще одной неотмеченной точкой. Предполагается, что неотмеченная точка, дающая максимальную норму, несет больше информации о данных в целом. В этой работе исследуются различные варианты этого критерия.

Ключевые слова: активное обучение, гауссовские процессы

1 Введение

В последнее время, возрастает интерес к активному обучению [6] – новому подходу к обучению, который позволит сократить размеры обучающей выборки. Недавно был предложен новый критерий для алгоритма активного обучения [1], разработанный в рассмотрении нейронных сетей с большим количеством параметров. Под последними имеются в виду такие модели, что отношения числа параметров p к размеру обучающей выборки n стремится к константе, большей единицы, при стремлении p и n к бесконечности. В этой работе исследуются различные варианты этого алгоритма, с той разницей, что вместо нейронных сетей ипользуются гауссовские процессы. Интуитивно, их связь можно объяснить следующим образом: выход слоя нейронной сети – есть линейная комбинация входов. Если считать признаки независимыми, а количество нейронов в слое достаточно большим, то имеет место центральная предельная теорема [4], которая утверждает, что сумма достаточно большого количества слабо зависимых случайных величин, имеет

распределение, близкое к нормальному. Гауссовский процесс же – бесконечно число случайных величин, любая конечная комбинация которых имеет нормальное распределение. Таким образом, есть связь между гауссовскими процессами и нейронными сетями с большим количеством параметров. В этой работе мы сосредоточимся на гауссовских процессах, а в будущем планируем заняться нейронными сетями.

2 Гауссовские процессы и их применение к задаче классификации

Здесь мы кратко введем некоторые необходимые для работы с ними определения. Полный обзор гауссовских процессов и их свойств представлен в [2].

Определение 1. *Гауссовский процесс $f(x)$ – стохастический процесс (совокупность случайных величин, индексированных некоторым параметром, чаще всего временем или координатами), такой что любой конечный набор этих случайных величин имеет многомерное нормальное распределение.*

Здесь в нашем случае $x \in \mathbb{R}^D$, где D - произвольное натуральное число. Гауссовский процесс полностью определяется функцией среднего и ковариационной функцией. Часто для упрощения записи функция среднего выбирается константой, равной 0. Введем обозначения:

$$m(x) = \mathbb{E}[f(x)] = 0, \tag{1}$$

$$k(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))]. \tag{2}$$

2.1 Регрессия

В задачах машинного обучения нам хочется получить апостериорное распределение на гауссовский процесс $f(x)$, пронаблюдав выборку данных X. Априорное распределение задается следующим образом:

$$\begin{pmatrix} f \\ f_* \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} K(X,X) & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{pmatrix}\right), \tag{3}$$

где $*$ соответствует тестовому множеству. С помощью формулы Байеса и некоторых алгебраических действий можно показать, что апостериорное распределение для незашумленных данных будет иметь следующий вид:

$$f_*|X, X_*, f \sim \mathcal{N}(\hat{f}, \hat{\sigma}^2), \tag{4}$$

где $\hat{f} = K(X_*, X)K(X, X)^{-1}f$ – среднее,
$\hat{\sigma}^2 = K(X_*, X_*) - K(X_*, X)K(X, X)^{-1}K(X, X_*))$ – дисперсия. Можно заметить, что выражение для дисперсии апостериорного распределения допускает интуитивно понятную интерпретацию: $K(X_*, X_*)$ – априорная ковариация и из нее вычитается положительное слагоемое, соответствующее

информации, взятой из обучающей выборки.
Таким образом, в случае регрессии апостериорное распределение дает нам среднее – предсказание гауссовского процессса в данной точке, а так же дисперсию – степень неуверенности модели в своем ответе. Это свойство дает гауссовским процессам преимущество перед другими моделями, хотя стоит заметить, что обращение матрицы в формуле (4) существенно замедляет вычисления.

2.2 Классификация

Ранее мы неявно предполагали, что правдоподобие $p(y|f, X)$ распределено нормально. Теперь же нельзя применять гауссовское правдоподобие, т.к. y – дискретная величина (класс, к которому принадлежит точка) – и это делает недоступным аналитическое выражение для апостериорного распределения. Рассмотрим для простоты задачу бинарной классификации. Тогда проблему действительных выводов легко решить, пропустив вывод гауссовского процесса через логистическую функцию $\sigma(f)$, которая "сжимает"все пространство $\mathbb{R}$ в отрезок $[0, 1]$. Сам гауссовский процесс – функция f – играет здесь роль латентной функции.
Вывод для предсказания разделяется в таком случае на два шага: во-первых, вычисление апостериорного распределения на латентную переменную f_*:

$$p(f_*|X, y, x_*) = \int p(f_*|X, x_*, f)p(f|X, y)df, \tag{5}$$

где $p(f|X, y) = p(y|f)p(f|X)/p(y|X)$, во-вторых, использование апостериорного распределения для маргинализации предсказания по всем возможным значениям латентной переменной:

$$p(y_* = +1|X, y, x_*) = \int p(f_*|X, y, x_*)\sigma(f_*)df_*. \tag{6}$$

Правдоподобие в выражении (5) лишает возможности вычислить интеграл (5) аналитически. Поэтому были разработаны различные методы аппроксимации этих интегралов. В работе [3] рассмотрены 6 вариантов таких аппроксимаций, а также приведено их всестороннее сравнение. Здесь же мы подробно опишем один из них.

Лапласовская аппроксимация. В этом методе апостериорное распределение $p(f|X, y)$ приближается нормальным распределением $q(f|X, y)$. Параметры этого распределения вычисляются на основе разложения в ряд Тейлора $logp(f|X, y)$ до второго порядка в максимуме апостериорного распределения:

$$q(f|X, y) \sim \mathcal{N}(f|\hat{f}, A^{-1}), \tag{7}$$

где $\hat{f} = argmax_f p(f|X, y)$, $A = -\nabla\nabla logp(f|X, y)|_{f=\hat{f}}$ – гессиан логарифма апостериорного распределения, взятого со знаком минус.

Параметры (7) вычисляются с помощью итеративного метода Ньютона. Достоинством этого метода является высокая скорость работы. Недостатком же – тот факт, что в алгоритм ориентируется на моду апостериорного распределения, а не на его среднее, что может привести к отклонениям в предсказаниях или меньшей уверенности в них. Стоит заметить, что наиболее популярным является другой алгоритм аппроксимации – Expectation Propagation. Его оценки среднего и дисперсии (значит, и точность) наиболее близки к действительности, по сравнению с другими алгоритмами.

3 Задача активного обучения

Теперь рассмотрим основные идеи активного обучения и метод, предложенный в работе [1]. В основе активного обучения лежит гипотеза о том, что модель сможет давать лучшие результаты на меньших обучающих выборках, если дать ей возможность самой выбирать точки датасета. Цикл работы заключается в следующем: модель выбирает следующую точку, которую ей хотелось бы иметь с меткой, запрашивает ее метку у "оракула" и обрабатывает полученную информацию, затем снова повторяет весь цикл. Обзор активного обучения представлен в [6].
Один из самых простых методов основывается на уверенности модели в точках выборки [7]. Например, для бинарной классификации, точки для получения меток - это точки, предсказанные вероятности для которых близки к 0.5. К этому же типу относится выбор новой точки с помощью энтропийного критерия и (уже в задаче регрессии) с помощью предсказанной дисперсии. Другой метод [8] подразумевает наличие нескольких моделей, обученных на текущем тренировочном множестве. Больше всего информации даст запрос точки, в которой данные модели больше всего не согласны друг с другом.

Max-min критерий и различные типы score-функций. Совсем недавно в работе [1] был представлен новый критерий для активного обучения, который мы рассмотрим в этой части.
Пусть $\mathcal{L} = (x_1, ..., x_n)$ – тренировочное множество с метками (y_1, y_n), $\mathcal{U}$ – набор примеров, не имеющих метки. Пусть f – функция, интерполирующая примеры из тренировочного множества, с минимальной нормой (это условие получено из опытов [10] – такие модели, как правило, имеют хорошие обощающие свойства). Определим $f_t^u(x)$ - как функцию, интерполирующую примеры из тренировочного множества, объединенного с точкой $u \in \mathcal{U}$ с меткой t, с минимальной нормой. Метку $t(u)$ будем выбирать одним из следующих способов:

$$t^{(1)}(u) = argmin_{t \in \{-1,1\}} \|f_t^u(x)\|,\ t^{(2)}(u) = \begin{cases} +1 \text{ if } f(u) \geq 0, \\ -1 \text{ if } f(u) < 0 \end{cases} \tag{8}$$

Определив $t(u)$, положим $f^u(x) = f_{t(u)}^u(x)$. Введем так же *score*-функции:

$$score^{(1)}(u) = \|f^u(x)\|,\ score^{(2)}(u) = \|f^u(x) - f(x)\|, \tag{9}$$

В первом случае больший *score* получит менее гладкая функция, во втором – та функция, которая наиболее сильно отличается от предыдущей интерполяции. Ожидается, что точка с большим *score*-ом наиболее информативная. Тогда следующая точка для получения метки есть

$$u^* = argmax_{u \in \mathcal{U}} score(u).$$

В выражениях (8) и (9) намеренно не определена конкретная норма для свободы в выборе различных вариаций *score*-функций. Если эти нормы выбраны одинаковы и выбрана первая *score*-функция, то новая точка u^* выбирается такой, что

$$\|f^{u^*}(x)\| = max_{u \in \mathcal{U}} min_{t \in \{-1,1\}} \|f_t^u(x)\|. \tag{10}$$

4 Эксперименты

Мы провели ряд экспериментов, в каждом из которых сравнивались 5 способов выбора следующей точки для получения ею метки (в скобках указано краткое обозначение):

1. *Случайный* – использовался для проверки адекватности полученных данных (rand).

2. *Дисперсия* – новая точка выбирается по максимуму дисперсии предсказания (vari).

3. *2-норма* – максимум $score^{(2)}(u) = \|f^u(x) - f(x)\|_{\mathcal{U}} = \sqrt{\sum\limits_{v \in \mathcal{U}} (f^u(v) - f(v))^2}$ по еще не отмеченным точкам (sqsm).

4. *RKHS-норма* – эта норма представлена в работе [1] (RKHS). Let's inference the $score^{(1)}(u)$.

   Using Schur complement formula we have $\widetilde{K}^{-1}$ as:
   $\|f_t^u(x)\| = \widetilde{\boldsymbol{y}}_{\boldsymbol{t}}^{\boldsymbol{T}} \widetilde{K} \widetilde{\boldsymbol{y}}_{\boldsymbol{t}}$
   $\widetilde{K}^{-1} = \begin{pmatrix} K^{-1} + K^{-1}\boldsymbol{a}(b - \boldsymbol{a}^T K^{-1}\boldsymbol{a})^{-1}\boldsymbol{a}^T K^{-1} & -K^{-1}\boldsymbol{a}(b - \boldsymbol{a}^T K^{-1}\boldsymbol{a})^{-1} \\ -(b - \boldsymbol{a}^T K^{-1}\boldsymbol{a})^{-1}\boldsymbol{a}^T K^{-1} & (b - \boldsymbol{a}^T K^{-1}\boldsymbol{a})^{-1} \end{pmatrix} =$
   $\begin{pmatrix} K^{-1} + \frac{K^{-1}\boldsymbol{a}\boldsymbol{a}^T K^{-1}}{b - \boldsymbol{a}^T K^{-1}\boldsymbol{a}} & \frac{-K^{-1}\boldsymbol{a}}{b - \boldsymbol{a}^T K^{-1}\boldsymbol{a}} \\ \frac{-\boldsymbol{a}^T K^{-1}}{b - \boldsymbol{a}^T K^{-1}\boldsymbol{a}} & \frac{1}{b - \boldsymbol{a}^T K^{-1}\boldsymbol{a}} \end{pmatrix}$

Then $\|f_t^u(x)\|$ turns to:

$$\|f_t^u(x)\| = \boldsymbol{y}^T K^{-1}\boldsymbol{y} + \boldsymbol{y}^T \frac{K^{-1}\boldsymbol{a}\boldsymbol{a}^T K^{-1}}{b - \boldsymbol{a}^T K^{-1}\boldsymbol{a}}\boldsymbol{y} - 2t\boldsymbol{y}^T K^{-1}\frac{\boldsymbol{a}}{b} -$$

$$-2t\boldsymbol{y}^T \frac{K^{-1}\boldsymbol{a}\boldsymbol{a}^T K^{-1}}{b - \boldsymbol{a}^T K^{-1}\boldsymbol{a}}\frac{\boldsymbol{a}}{b} + \frac{1}{b - \boldsymbol{a}^T K^{-1}\boldsymbol{a}} =$$

$$= \boldsymbol{y}^T K^{-1}\boldsymbol{y} + \frac{\left(\boldsymbol{y}^T K^{-1}\boldsymbol{a}\right)^2}{b - \boldsymbol{a}^T K^{-1}\boldsymbol{a}} - 2t\frac{\boldsymbol{y}^T K^{-1}\boldsymbol{a}}{b}\left(1 + \frac{\boldsymbol{a}^T K^{-1}\boldsymbol{a}}{b - \boldsymbol{a}^T K^{-1}\boldsymbol{a}}\right) + \frac{t^2}{b - \boldsymbol{a}^T K^{-1}\boldsymbol{a}} =$$

$$= \boldsymbol{y}^T K^{-1}\boldsymbol{y} + \frac{\left(\boldsymbol{y}^T K^{-1}\boldsymbol{a}\right)^2}{b - \boldsymbol{a}^T K^{-1}\boldsymbol{a}} - \frac{2t\left(\boldsymbol{y}K^{-1}\boldsymbol{a}\right)}{b - \boldsymbol{a}^T K^{-1}\boldsymbol{a}} + \frac{t^2}{b - \boldsymbol{a}^T K^{-1}\boldsymbol{a}} =$$

$$= \boldsymbol{y}^T K^{-1}\boldsymbol{y} + \frac{\left(\boldsymbol{y}^T K^{-1}\boldsymbol{a} - t\right)^2}{b - \boldsymbol{a}^T K^{-1}\boldsymbol{a}} = \boldsymbol{y}^T K^{-1}\boldsymbol{y} + \frac{\left(f(u) - y(u)\right)^2}{b - \boldsymbol{a}^T K^{-1}\boldsymbol{a}} \tag{11}$$

$$score^{(1)}(u) = \|f_t^u(x)\| = \|f(x)\|_{\mathcal{H}} + \frac{(1 - t\cdot f(u))^2}{b - \boldsymbol{a^T}\cdot\boldsymbol{K^{-1}}\cdot\boldsymbol{a}}, \tag{12}$$

где $b = K(u,u), \boldsymbol{a} = [K(x_1,u), ..., K(x_n,u)]^T$. Важно заметить, что знаменатель дроби в данном выражении - есть дисперсия предсказания модели в точке u.

5. *RKHS-норма · дисперсия* – оказывается, RKHS-норма ведет себя довольно нестабильно. Умножение на дисперсию не только позволяет решить эту проблему, но и улучшить качество работы модели - этот ход был предложен в работе [9] (Hvar).

$$(\|f_t^u(x)\| - \|f(x)\|_{\mathcal{H}}) \cdot \mathsf{D} = (1 - t\cdot f(u))^2.$$

Let $\widetilde{a} = \begin{pmatrix} \boldsymbol{a} \\ K(u,x) \end{pmatrix}$. Where u is the new point in out dataset and x is the point we calculate $f(x)$ for. Let's compute difference $f_u(x) - f(x)$:
$f_u(x) - f(x) = \widetilde{\boldsymbol{a}}^T \widetilde{K}^{-1}\widetilde{\boldsymbol{y}} - \boldsymbol{a}^T K^{-1}\boldsymbol{y} =$
$\left(\boldsymbol{a}^T\ K(u,x)\right)\begin{pmatrix} K^{-1} + \frac{K^{-1}\boldsymbol{a}\boldsymbol{a}^T K^{-1}}{b - \boldsymbol{a}^T K^{-1}\boldsymbol{a}} & \frac{-K^{-1}\boldsymbol{a}}{b - \boldsymbol{a}^T K^{-1}\boldsymbol{a}} \\ \frac{-\boldsymbol{a}^T K^{-1}}{b - \boldsymbol{a}^T K^{-1}\boldsymbol{a}} & \frac{1}{b - \boldsymbol{a}^T K^{-1}\boldsymbol{a}} \end{pmatrix}\begin{pmatrix} \boldsymbol{y} \\ t \end{pmatrix} - \boldsymbol{a}^T K^{-1}\boldsymbol{y} =$
$\frac{(\boldsymbol{a}^T K^{-1}\boldsymbol{a} - K(u,x))(\boldsymbol{a}^T K^{-1}\boldsymbol{y} - t)}{b - \boldsymbol{a}^T K^{-1}\boldsymbol{a}}$
Заметим, что первые два способа не требуют определения t и *score*-функции, а в последнем *score*-функция модифицирована и не соответсвует какой-либо стандартной норме. В экспериментах был использован второй способ выбора метки t (см. (8)).

Одномерные данные Из многомерного нормального распределения с нулевым средним и матрицей ковариации, заданной ядром RBF $cov(x,x') = \sigma^2 \cdot \exp - \frac{\|x - x'\|^2}{2}$, были сгенерированы 1000 точек – знчения гауссовского процесса (латентной переменной) – $f(x)$. Затем эти значения были пропущены через логистическую функцию. Эти точки сопоставились 1000 точек

из отрезка $[0, 1]$ – x. Полученные значения интерпретировались как вероятности, и по ним точки x были разделены на два класса.
Размер отложенной выборки – 350 точек. Обучающая выборка в начале имела 10 точек и на каждой итерации дополнялась одной точкой, выбранной согласно одному из критериев. При учете новой точки u параметры гауссовского процесса – функции среднего и ковариации – не менялись относительно предыдущего процесса, обученного без точки u из предположения, что всего одна точка в обучающей выборке не может значительно изменить эти параметры. Однако, каждые 20 итераций они оптимизировались отдельно от добавления новых точек. Этот алгоритм повторялся до тех пор, пока в обучающей выборке не окажутся все 650 точек.
На рис. 1 построена зависимость точности классификации на отложенной выборке от размера обучающей выборки. Для наглядности график ограничен размером обучающей выборки – 80. Как видно из графиков, лучше всего на этих данных работают методы sqsm и Hvar.
На рис. 2 изображены графики *score*-функций на одной из итерации алго-

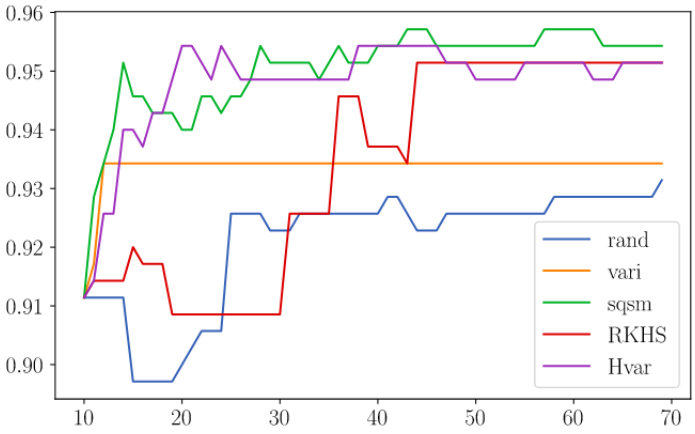

Рис. 1: Графики точности классификации на отложенной выборке в зависимости от размера обучающей выборки для одномерных данных.

ритма активного обучения для этих методов. На графики так же добавлены вероятности принадлежности точек x к классу 1. Это позволяет заметить, что локальные максимумы расположены близко к границам разделения классов.

Многомерные синтетические данные Для задачи бинарной классификации были сгенерированы 5-мерные данные. Аналогичным образом построены графики точности классификации в зависимости от размера обучающей выборки, представленные на рис. 3. Здесь лучше всего работают методы RKHS и Hvar.

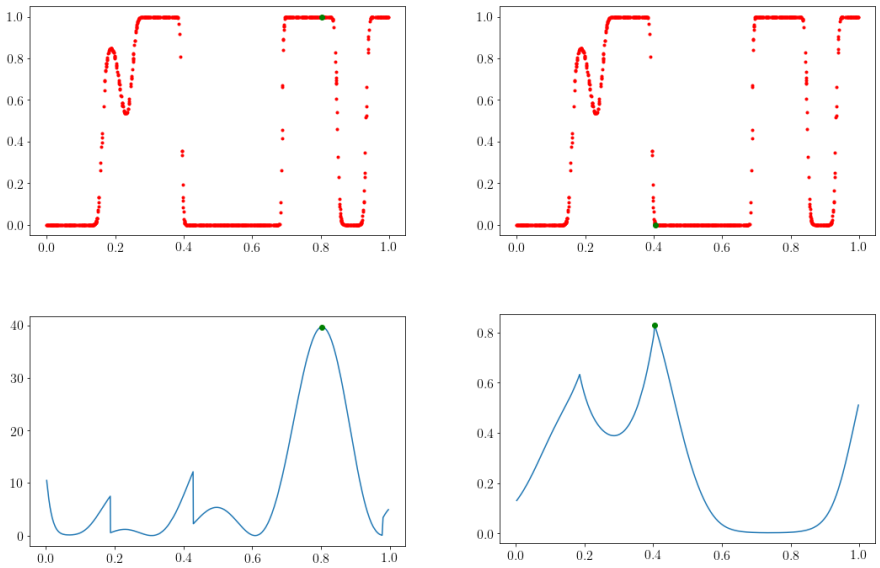

Рис. 2: Графики *score*-функций на одной из итерации алгоритма активного обучения. На верхних рисунках изображены вероятности принадлежности точек на оси x к классу 1. Зеленые точки обозначают точку с максимальным значением *score*-функции. Нижние рисунки иллюстрируют значения *score*-функций, соответствующие двум методам: Hvar (слева) и sqsm (справа).

5 Заключение

В этой работе мы рассмотрели применение гауссовских процессов к задаче классификации и различные варианты алгоритмов активного обучения и сравнили их качество. Нам так же удалось посмотреть на их работу "изнутри" с помощью графиков *score*-функций. По имеющимся результатам можно сказать лишь, что в условиях, когда данных мало, хорошо работают методы sqsm, RKHS и Hvar. Причем последний хорошо показывает себя в обоих рассмотренных задачах. Первый же имеет наиболее наглядную интерпретацию в плане графиков *score*-функций.
Заметим, что ранее таким сравнением методов никто не занимался. В дальнейшем, безусловно, планируется поставить более масштабные эксперименты, т.к. небольшой размер тестовой выборки вызывает дискретность значний точности, из-за чего графики трудно воспринимать.

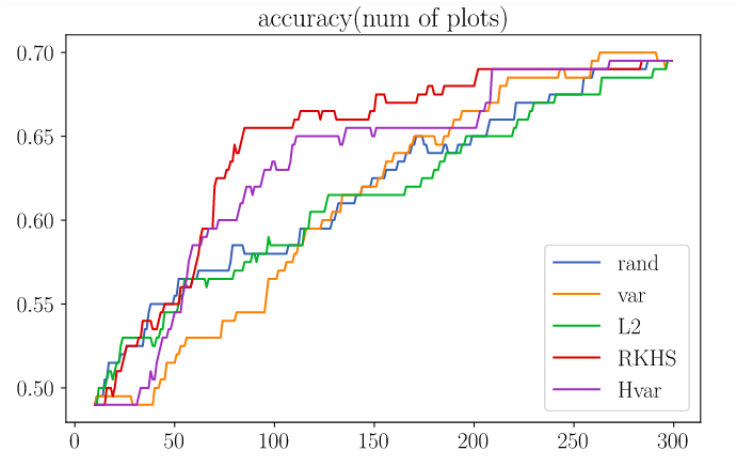

Рис. 3: Графики точности классификации на отложенной выборке в зависимости от размера обучающей выборки для 5-мерных данных.

Список литературы

[1] Mina Karzand, Robert D. Nowak: Active Learning in the Overparameterized and Interpolating Regime arXiv preprint arXiv:1905.12782,2019

[2] C. E. Rasmussen, C. K. I. Williams, Gaussian Processes for Machine Learning, the MIT Press, 2006, ISBN 026218253X

[3] Hannes Nickisch, Carl Edward Rasmussen, Approximations for Binary Gaussian Process Classification, Journal of Machine Learning Research 9 (2008) 2035-2078

[4] Гнеденко Б. В., Курс теории вероятностей, УРСС. М.: 2001

[5] Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. In International Conference on Machine Learning, pages 3331–3340, 2018.

[6] Burr Settles. Active Learning Literature Survey. 2010.

[7] D. Lewis and W. Gale. A sequential algorithm for training text classifiers. In Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, pages 3–12. ACM/Springer, 1994.

[8] H.S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In Proceedings of the ACM Workshop on Computational Learning Theory, pages 287–294, 1992

[9] Burnaev E., Panov M. (2015) Adaptive Design of Experiments Based on Gaussian Processes. In: Gammerman A., Vovk V., Papadopoulos H. (eds) Statistical Learning and Data Sciences. SLDS 2015. Lecture Notes in Computer Science, vol 9047. Springer, Cham

[10] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. arXiv preprint arXiv:1903.08560, 2019.