Байесовское активное обучение для классификации и изучения предпочтений

Нил Хоулсби, Ференц Хушар, Зубин Гахрамани, Мате Ленгиел
Лаборатория вычислительного и биологического обучения
Кембриджский университет 30

декабря 2011 г.

Аннотация

Информационно-теоретическое активное обучение широко изучается для вероятностных моделей. Для простой регрессии оптимальная близорукая политика легко поддается решению. Однако для других задач и более сложных моделей, таких как классификация с непараметрическими моделями, оптимальное решение вычислить сложнее. Существующие подходы делают приближения, чтобы достичь трактибильности. Мы предлагаем подход, выражающий информационный выигрыш в терминах предсказательных энтропий, и применяем этот метод к классификатору гауссовских процессов (GPC). Наш подход делает минимальные приближения к полной информационно-теоретической цели. Наши экспериментальные результаты выгодно отличаются от многих популярных алгоритмов активного обучения и имеют равную или меньшую вычислительную сложность. Мы также хорошо сравниваемся с теоретическими подходами, которые имеют больше информации и требуют гораздо больше вычислительного времени. Во-вторых, развивая переформулировку бинарного обучения предпочтениям в задачу классификации, мы расширяем наш алгоритм до обучения предпочтениям с помощью гауссовского процесса.

1 Введение

В большинстве систем машинного обучения обучаемый пассивно собирает данные, на основе которых он делает выводы о своем окружении. Однако при активном обучении обучаемый ищет наиболее полезные измерения для обучения. Цель активного обучения - создать наилучшую модель при минимальном количестве данных; это тесно связано со статистической областью оптимального экспериментального проектирования. С появлением Интернета и расширением хранилищ стало доступно огромное количество немаркированных данных, но получение меток может быть дорогостоящим. Для поиска наиболее полезных данных в этом огромном пространстве требуются эффективные алгоритмы активного обучения.

Два подхода к активному обучению - это использование теории решений и теории информации [Kapoor et al.] В первом случае минимизируется ожидаемая

потери, возникающие после принятия решений на основе собранных данных, т. е. минимизировать апостериорный риск Байеса [Roy and McCallum, 2001]. Максимизация производительности при тестировании является конечной целью большинства обучающихся, однако оценить эту цель может быть очень сложно. Например, методы, предложенные в [Kapoor et al., 2007, Zhu et al., 2003] для классификации, в общем случае являются дорогостоящими для вычисления. Кроме того, можно не знать заранее функцию потерь или тестовое распределение, а также хотеть, чтобы модель хорошо работала при различных функциях потерь. В экстремальных сценариях, таких как анализ исследовательских данных или визуализация, потери может быть очень сложно оценить количественно.

Это мотивирует информационно-теоретические подходы к активному обучению, которые не зависят от решаемой задачи и конкретных тестовых данных, что известно как индуктивный подход. Они стремятся как можно быстрее сократить число выполнимых моделей, используя либо эвристику (например, маржинальную выборку [Tong and Koller, 2001]), либо формализацию неопределенности с помощью хорошо изученных величин, таких как энтропия Шеннонса и KL-расхождение [Cover et al., 1991]. Хотя последний подход был предложен несколько десятилетий назад [Lindley, 1956, Bernardo, 1979], не всегда просто применить критерии к сложным моделям, таким как непараметрические процессы с бесконечным пространством параметров. В результате существует множество алгоритмов, которые вычисляют приближенные апостериорные энтропии, выполняют выборку или работают со смежными величинами в неверояностных моделях.

Мы возвращаемся к этой проблеме, представляем полный информационный критерий и демонстрируем, как применить его к классификации гауссовских процессов (GPC), получая новый алгоритм активного обучения, который делает минимальные приближения. GPC - это мощная непараметрическая модель на основе ядра, представляющая собой интересную проблему для информационно-теоретического активного обучения, поскольку пространство параметров имеет бесконечную размерность, а апостериорное распределение аналитически неразрешимо. В разделе 2 мы представляем информационно-теоретический подход к активному обучению. В разделе 3 мы применяем его к GPC и показываем, как распространить наш метод на обучение предпочтениям. В разделе 4 мы рассматриваем другие подходы и их сравнение с нашим алгоритмом. Особое внимание мы уделяем сравнению нашего подхода с Информативной Векторной Машиной, которая непосредственно занимается выбором точек данных для GP. В разделе 5 мы представляем результаты на различных наборах данных, а в разделе 6 делаем выводы.

2 Байесовская информационная теория активного обучения

Мы рассматриваем полностью дискриминативную модель, в которой целью активного обучения является обнаружение зависимости некоторой переменной $y \in Y$ от входной переменной $\boldsymbol{x}$. Ключевая идея активного обучения заключается в том, что обучаемый выбирает входные запросы $\boldsymbol{x}_i \in X$ и наблюдает за реакцией системы y_i , а не пассивно получает пары ($\boldsymbol{x}\ \boldsymbol{y}_{ii}$).

В рамках байесовского подхода мы предполагаем существование некоторых скрытых параметров $\boldsymbol{\theta}$, которые управляют зависимостью между входами и выходами, $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$.

Имея наблюдаемые данные $D = \{(\boldsymbol{x}_i , y_{i=1})\}^n$, можно построить апостериорное распределение по па-

раметров выводится значение $p(\theta \mid D)$. Главная цель информационно-теоретического обучения - максимально быстрое сокращение числа возможных гипотез, то есть минимизация неопределенности относительно параметров с помощью энтропии Шеннона [Cover et al., 1991]. Выбираются точки данных D', которые удовлетворяют $\arg\min_{D'} H[\theta \mid D'] = -\int p(\theta \mid D') \log p(\theta \mid D') d\theta$. Решение этой задачи в общем случае является NP-трудным, однако, как обычно бывает в задачах последовательного принятия решений, делается близорукое (жадное) приближение [Heckerman et al.] Было показано, что миопическая политика может быть близка к оптимальной [Golovin and Krause, 2010, Dasgupta, 2005]. Таким образом, цель состоит в поиске точки данных $\boldsymbol{x}$, которая максимизирует уменьшение ожидаемой апостериорной энтропии:

$$\arg\max_{\boldsymbol{x}} H[\theta \mid D] - E_{y \sim p(y \mid \boldsymbol{x} D)} [H[\theta \mid y, \boldsymbol{x}, D]]. \quad (1)$$

Обратите внимание, что требуется ожидание по неизвестному выходу y. Многие работы Например, в работах [MacKay, 1992, Krishnapuram et al., , Lawrence et al., 2003] предлагается использовать эту цель напрямую. Однако постеры параметров часто имеют большую размерность, и вычисление их энтропий обычно трудновыполнимо. Более того, для непараметрических процессов пространство параметров имеет бесконечную размерность, поэтому уравнение (1) становится плохо определенным. Чтобы избежать сетки пространства параметров (экспоненциально сложной с ростом размерности) или выборки (по которой, как известно, трудно оценить энтропию без внесения погрешности [Panzeri and Petersen, 2007]), в этих работах делаются гауссовские или низкоразмерные аппроксимации и вычисляется энтропия приближенного апостериора. Возникает вторая вычислительная трудность: если рассматривается $N_{\boldsymbol{x}}$ точек данных и может быть замечено N_y ответов, то для вычисления уравнения (1) требуется ($N\ N_{\boldsymbol{x}y}$) потенциально дорогих обновлений апостериорных данных.

Важное понимание возникает, если мы заметим, что цель в уравнении (1) эквивалентна условной взаимной информации между неизвестным выходом и параметрами, $I[\boldsymbol{\theta}, y \mid \boldsymbol{x}, D]$. Используя это понимание, легко показать, что задача может быть перестроена для вычисления энтропии в пространстве y:

$$\arg\max_{\boldsymbol{x}} H[\boldsymbol{y} \mid \boldsymbol{x}, D] - E_{\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid D)} [H[\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}]]. \quad (2)$$

Уравнение (2) преодолевает трудности, описанные нами для уравнения (1). Энтропии теперь вычисляются в обычно низкоразмерном пространстве выходов. Для бинарной классификации это просто энтропии переменных Бернулли. Также $\boldsymbol{\theta}$ теперь зависит только от D, поэтому требуется только O (1) обновление апостериорных данных. Уравнение (2) также дает нам интересную интуицию относительно цели; мы ищем $\boldsymbol{x}$, для которого модель имеет минимальную неопределенность относительно y (высокая $H[\boldsymbol{y} \mid \boldsymbol{x}, D]$), но для которого индивидуальные настройки параметров являются уверенными (низкая $E_{\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid D)} [H[y \mid \boldsymbol{x}, \boldsymbol{\theta}]]$). Это можно интерпретировать как поиск $\boldsymbol{x}$, для которого параметры в апостериоре больше всего расходятся во мнениях о результате, поэтому мы называем эту задачу "Байесовское активное обучение по расхождениям" (BALD). Мы представляем метод, позволяющий применить уравнение (2) непосредственно к GPC и обучению предпочтениям. Нам больше не нужно строить вычисление энтропии в зависимости от типа апостериорной аппроксимации (как

в [MacKay, 1992, Krishnapuram et al., , Lawrence et al., 2003]), но могут свободно выбирать из многих доступных алгоритмов. При этом вводятся минимальные дополнительные приближения, и поэтому, насколько нам известно, наш алгоритм представляет собой наиболее точный и быстрый способ выполнения полного информационно-теоретического активного обучения в непараметрических дискриминантных моделях.

3 Гауссовские процессы для классификации и предварительного обучения

В этом разделе мы выводим алгоритм BALD для классификации на основе гауссовых процессов (GPC). ГП - мощный и популярный непараметрический инструмент для регрессии и классификации. GPC представляется особенно сложной задачей для информационно-теоретического активного обучения, поскольку пространство параметров бесконечно, однако, используя (2), мы можем полностью вычислить соответствующие информационные величины без необходимости вычислять энтропии объектов бесконечной размерности. Вероятностная модель, лежащая в основе GPC, выглядит следующим образом:

$$f \sim \mathrm{GP}(\mu(\cdot), k(\cdot,\cdot))$$
$$\boldsymbol{y}|\boldsymbol{x}, f \sim \mathrm{Bernoulli}(\Phi(f(\boldsymbol{x})))$$

Латентный параметр, который теперь называется f, является функцией $\boldsymbol{x}$ R, и ему присваивается приоритет гауссовского процесса со средним $\mu(\cdot)$ и ковариационной функцией или ядром $k(\cdot,\cdot)$. Мы рассматриваем случай пробита, когда при значении f , y принимает распределение Бернулли с вероятностью $\Phi(f(\boldsymbol{x}))$, а Φ - гауссова CDF. Более подробно о ГП см. в [Rasmussen and Williams, 2005].

Вывод в модели GPC является трудноразрешимой задачей; при некоторых $\boldsymbol{D}$ наблюдениях апостериор над f становится негауссовым и сложным. Наиболее часто используемые методы приближенного вывода - EP, аппроксимация Лапласа, фильтрация по предполагаемой плотности и разреженные методы - все аппроксимируют апостериор гауссовкой [Rasmussen and Williams, 2005]. На протяжении всего этого раздела мы будем считать, что нам предоставлена такая гауссова аппроксимация одним из этих методов, хотя алгоритму активного обучения все равно, каким именно. В нашем выводе мы будем использовать $\approx^{1}$, чтобы указать, где такая аппроксимация является эксплуатируется.

Информативность запроса $\boldsymbol{x}$ вычисляется с помощью уравнения (2). Энтропия бинарной выходной переменной y при фиксированном значении f может быть выражена в терминах бинарной энтропийной функции h:

$$\mathrm{H}[\boldsymbol{y}|\boldsymbol{x}, f] = \mathrm{h}(\Phi(f(\boldsymbol{x})))$$
$$\mathrm{h}(p) = -p \log p - (1 - p) \log(1 - p)$$

Необходимо вычислить ожидания по апостериору. Используя гауссово приближение к апостериору, для каждого $\boldsymbol{x}$, $f_{\boldsymbol{x}} = f(\boldsymbol{x})$ будет следовать гауссову распределению со средним значением $\mu_{\boldsymbol{x},D}$ и $_{\boldsymbol{x},D}$. Для вычисления уравнения (2) необходимо вычислить дисперсией σ^2

две энтропийные величины. Первый член в уравнении (2), $H[y \mid \mathbf{x}, D]$, может быть обработан аналитически для случая пробит:

$$H[y \mid \mathbf{x}, D] = h \int \Phi(f_{\mathbf{x}}) N(f_{\mathbf{x}} \mid \mu_{\mathbf{x},D}, \sigma^2_{\mathbf{x},D}) df_{\mathbf{x}} = h\left(\Phi\left(\frac{\mu_{\mathbf{x},D}}{\sqrt{\sigma^2_{\mathbf{x},D} + 1}}\right)\right) \tag{3}$$

Второй член, $E_{f \sim p(f \mid D)}[H[y \mid \mathbf{x}, f]]$, может быть вычислен приблизительно следующим образом:

$$\begin{aligned} & E_{f \sim p(f \mid D)}[H[y \mid \mathbf{x}, f]] \\ & \overset{1}{\approx} \int h(\Phi(f_{\mathbf{x}})) N(f_{\mathbf{x}} \mid \mu_{\mathbf{x},D}, \sigma^2_{\mathbf{x},D}) df_{\mathbf{x}} \\ & \overset{2}{\approx} \text{эксп}\left(-\frac{f_{\mathbf{x}}^2}{\pi \ln 2}\right) N(f_{\mathbf{x}} \mid \mu_{\mathbf{x},D}, \sigma^2_{\mathbf{x},D}) df_{\mathbf{x}} \\ & = \frac{C}{\sqrt{\sigma^2_{\mathbf{x},D} + C^2}} \exp\left(-\frac{\mu^2_{\mathbf{x},D}}{2(\sigma^2_{\mathbf{x},D} + C^2)}\right) \end{aligned} \tag{4}$$

где $C = \sqrt{\frac{\pi \ln 2}{2}}$. Первое приближение, $\overset{1}{\approx}$, отражает гауссову ап-. апроксимация к апостериору. Интеграл в левой части уравнения (4) является трудновыполнимым. Выполнив разложение Тейлора по $\ln h(\Phi(f_{\mathbf{x}}))$ (см. дополнительный материал), мы видим, что он может быть аппроксимирован до $O(f_{\mathbf{x}}^4)$ квадратичной экспоненциальной кривой, $\exp(-f_{\mathbf{x}}^2 / \pi \ln 2)$. Мы будем называть это приближение $\overset{2}{\approx}$. Теперь мы можем применить стандартную формулу свертки для гауссианов, чтобы получить выражение в замкнутой форме для обоих членов уравнения (2).

На рис. 1 показана поразительная точность этого простого приближения. Максимально возможная ошибка при использовании этого приближения будет иметь место, если $N(f_{\mathbf{x}} \mid \mu_{\mathbf{x},D}, \sigma^2_{\mathbf{x},D})$ центрируется на $\mu_{\mathbf{x},D} = \pm 2{,}05$, при этом $\sigma^2_{\mathbf{x},D}$ стремится к нулю (см. Рис. 1, абсолютная погрешность), давая лишь 0,27% погрешности в интеграле в уравнении (4). Авторам неизвестно о предыдущем использовании этого простого и полезного приближения в данном контексте. В разделе 5 мы экспериментально исследуем потерю информации из приближений $\overset{1}{\approx}$ и $\overset{2}{\approx}$ по сравнению с золотым стандартом обширного Моделирование методом Монте-Карло.

Вкратце, алгоритм BALD для классификации гауссовских процессов состоит из двух шагов. Сначала применяется любой стандартный алгоритм приближенного вывода для GPC (например, EP) для получения апостериорных прогнозных средних $\mu_{\mathbf{x},D}$ и $\sigma_{\mathbf{x},D}$ для каждой интересующей точки $\mathbf{x}$. Затем выбирается запрос $\mathbf{x}$, который максимизирует следующую объективную функцию:

$$h\left(\Phi\left(\frac{\mu_{\mathbf{x},D}}{\sqrt{\sigma^2_{\mathbf{x},D} + 1}}\right)\right) - \frac{C \exp\left(-\frac{\mu^2_{\mathbf{x},D}}{2(\sigma^2_{\mathbf{x},D} + C^2)}\right)}{\sqrt{\sigma^2_{\mathbf{x},D} + C^2}} \tag{5}$$

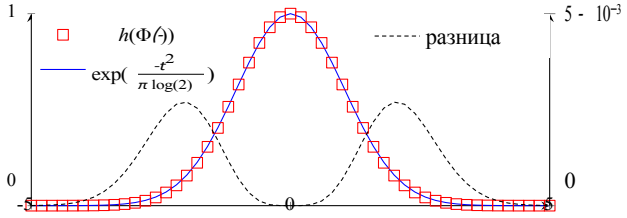

Рисунок 1. Аналитическая аппроксимация ( ) бинарной энтропии функции ошибки ( ) квадратичной экспонентой ( ). Абсолютная ошибка ( ) не превышает 3 - 10 .

Для большинства практически значимых ядер цель (5) является гладкой и дифференцируемой функцией от $\boldsymbol{x}$, поэтому для поиска максимально информативного запроса можно использовать процедуры оптимизации на основе градиента.

3.1 Расширение: Изучение гиперпараметров

Во многих приложениях набор параметров $\boldsymbol{\theta}$ естественным образом делится на интересующие параметры $\boldsymbol{\theta}^+$ и на параметры, не являющиеся существенными, $\boldsymbol{\theta}^-$, т. е. $\boldsymbol{\theta} = \boldsymbol{\theta}^+, \boldsymbol{\theta}^-$. В таких условиях активное обучение может потребовать запросить точки, которые максимально информативны в отношении $\boldsymbol{\theta}^+$, не заботясь при этом о $\boldsymbol{\theta}^-$. Интегрируя уравнение (1) по параметрам помех, $\boldsymbol{\theta}^-$, цель BALD переформулируется как:

$$\mathrm{H}\,\mathrm{E}_{+\,p(\boldsymbol{\theta},\boldsymbol{\theta}^-|D)}\,[\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}^+,\boldsymbol{\theta}^-] - \mathrm{E}_{+\,\mathrm{H}_{p(\boldsymbol{\theta}|D)}}\,\mathrm{E}_{p(\boldsymbol{\theta}^-|\boldsymbol{\theta},D)}\,[\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}^+,\boldsymbol{\theta}^-] \qquad (6)$$

В контексте моделей GP гиперпараметры обычно управляют гладкостью или пространственным масштабом функций. Если мы сохраняем апостериорное распределение по этим гиперпараметрам, что можно сделать, например, с помощью гамильтонова метода Монте-Карло, то мы можем либо рассматривать их в качестве ненужных параметров $\boldsymbol{\theta}^-$ и использовать уравнение 6, либо включить их в $\boldsymbol{\theta}^+$ и проводить активное обучение и по ним. В некоторых случаях, например при автоматическом определении релевантности [Rasmussen and Williams, 2005], может даже иметь смысл рассматривать гиперпараметры как переменные, представляющие основной интерес, а саму функцию f - как параметр $\boldsymbol{\theta}^-$.

3.2 Изучение предпочтений

Наша система активного обучения для GPC может быть расширена на важную проблему обучения предпочтениям [Fürnkranz and Hüllermeier, 2003, Chu and Ghahramani, 2005]. При обучении предпочтениям набор данных состоит из пар элементов $(\boldsymbol{u}_i, \boldsymbol{v}_i) \in X^2$ с бинарными метками, $y_i \in \{0, 1\}$. $y_i = 1$ означает, что экземпляр $\boldsymbol{u}_i$ предпочтительнее $\boldsymbol{v}_i$, обозначается $\boldsymbol{u}_i \succ \boldsymbol{v}_i$. Задача состоит в том, чтобы предсказать отношение предпочтения между любыми $(\boldsymbol{u}, \boldsymbol{v})$. Мы можем рассматривать это как частный случай построения классификатора на парах входных данных

h :2 ℵ, 1→ В работе [Chu and Ghahramani, 2005] предлагается байесовский подход, использующий латентную функцию предпочтения f , для которой определяется GP-приор. Модель предсказывает предпочтение, $u_i > v_i$ всякий раз, когда $f(u_i) + \epsilon_{u_i} > f(v_i) + \epsilon_{v_i}$, где ϵ_{u_i} , ϵ_{v_i} обозначают аддитивный гауссовский шум. В рамках этой модели вероятность f становится равной:

$$P[y = 1 \mid (u_i, v_i), f] = P[u_i > v_i \mid f] = \Phi \frac{f(u_i) - f(v_i)}{\sqrt{2\sigma_{шум}}} \quad (7)$$

При изменении масштаба латентной функции f можно предположить, что $\sqrt{2\sigma_{noise}} = 1$. Вероятность зависит только от разницы между $f(u)$ и $f(v)$. Поэтому мы определяем $g(u, v) = f(u) f(v)$ и делаем вывод полностью в терминах g, для которого вероятность становится такой же, как и для пробит-классификации: $y \mid u, v, f \sim$ Bernoulli(Φ($g(u, v)$)). Мы замечаем, что GP-приор индуцирован на g, поскольку он формируется путем выполнения линейной операции над f , для которого у нас уже есть GP-приор $f \sim$ GP(0, k). Мы можем вывести индуцированную ковариационную функцию g (вывод в Дополнительных материалах) следующим образом: $k_{\text{pref}}((u_i, v_i), (u_j, v_j)) = k(u_i, u_j) + k(v_i, v_j) - k(u_i, v_j) - k(v_i, u_j)$).

Заметим, что это ядро k_{pref} соблюдает свойства антисимметрии, необходимые для сценария обучения предпочтениям, т. е. значение $g(u, v)$ идеально антикоррелировано с $g(v, u)$, что обеспечивает выполнение условия $P[u > v] = 1 - P[v > u]$. Таким образом, мы можем заключить, что GP-система обучения предпочтениям из [Chu and Ghahramani, 2005] эквивалентна GPC с определенным классом ядер, которые мы можем назвать *ядрами суждения о предпочтениях*. Таким образом, наш алгоритм активного обучения, представленный в разделе 3 для GPC, может быть легко применен и к парному обучению предпочтениям.

4 Смежные методологии

Существует ряд тесно связанных между собой алгоритмов активной классификации, которые мы сейчас и рассмотрим.

Информативная векторная машина (IVM): Возможно, наиболее близким подходом является IVM [Lawrence et al., 2003]. Этот популярный и успешный подход к активному обучению был разработан специально для GP; он использует информационно-теоретический подход и поэтому очень похож на BALD. Алгоритм IVM был разработан для подвыборки набора данных для обучения ГП, поэтому ему известны значения y до включения измерения; поэтому он не может работать явно в пространстве выходов, т. е. с уравнением (2). В IVM используется уравнение (1), но энтропии параметров вычисляются приблизительно в маргинальном подпространстве, соответствующем наблюдаемым точкам данных. Уменьшение энтропии после включения новой точки данных может быть эффективно вычислено с помощью ковариационной матрицы ГП.

Хотя IVM и BALD преследуют одну и ту же цель, они работают принципиально по-разному, когда выполняется приближенный вывод. В любой момент времени

Оба метода имеют приближенное апостериорное значение $q_t(\theta|D)$, которое может быть обновлено с учетом правдоподобия новой точки данных $p(y_{t+1}|f, \mathbf{x}_{t+1})$, что дает $\hat{p}_{t+1}(\theta|D, \mathbf{x}_{t+1}, y_{t+1}) = \frac{1}{Z} q_t(\theta|D)p(y_{t+1}|f, \mathbf{x}_{t+1})$. Если апостериор при $t+1$ аппроксимируется непосредственно одним
получает $q_{t+1}(\theta|D, \mathbf{x}_{t+1}, y_{t+1})$. BALD вычисляет разность энтропии между q_t и $\hat{p}_{t+1}$, без необходимости вычислять q_{t+1} для каждого кандидата $\mathbf{x}$. В отличие от этого, IVM вычисляет изменение энтропии между q_t и q_{t+1}. Ап-процесс IVM не может вычислить энтропию всего бесконечного постерного ряда, и требует $(N\ N_{\mathbf{x}y})$ обновлений постерного ряда. Для эффективного обновления апостериорной информации используется фильтрация по предполагаемой плотности (ADF). Использование ADF означает, что q_{t+1} является прямым приближением к $\hat{p}_{t+1}$, что указывает на то, что IVM делает дальнейшее приближение к BALD. Поскольку BALD требует только (1) обновления апостериорных данных, он может позволить себе использовать более точные итерационные процедуры, такие как EP.

Информационно-теоретические подходы: Выборка с максимальной энтропией (MES) [Sebastiani and Wynn, 2000] явно работает в пространстве данных (Eqn. (2)). MES была предложена для регрессионных моделей с независимым от входа шумом наблюдения. Хотя уравнение (2) и используется, второй член является постоянным из-за шума, не зависящего от входа, и игнорируется. Однако MES нельзя использовать для гетероскедастической регрессии или классификации: она не позволяет провести различие между неопределенностью модели и неопределенностью наблюдения (о которой наша модель может быть уверена). Некоторые игрушечные демонстрации показывают, что этот "основанный на информации" критерий активного обучения патологически работает в классификации, если многократно запрашивать точки вблизи границы принятия решения или в областях с высокой неопределенностью наблюдений, например [Huang et al., 2010]. Это происходит потому, что MES не подходит для этой области; BALD различает неопределенность наблюдения и модели и устраняет эти проблемы, к а к мы покажем.

Целевые функции, основанные на взаимной информации, представлены в [Ertin et al., Fuhrmann, 2003]. Они максимизируют взаимную информацию между измеряемой переменной и интересующей переменной. Фурманн [Fuhrmann, 2003] применяет это к линейным гауссовским моделям и акустическим массивам, а Эртин *и др* [Ertin et al. Хотя эти задачи и связаны между собой, они не работают с параметрами модели и не применяются для классификации. В задачах [Guestrin et al., 2005, Krause et al., 2006] также используется взаимная информация. Они заранее определяют точки интереса и максимизируют ожидаемую взаимную информацию между прогнозируемыми распределениями в этих точках и в наблюдаемых местах. Хотя эта задача перспективна для регрессии, она не подходит для моделей с шумом наблюдения, зависящим от входных данных, таких как классификация или обучение предпочтениям.

Теоретические решения: Мы кратко упомянем теоретико-решающие подходы к активному обучению. Два тесно связанных алгоритма, [Kapoor et al., 2007, Zhu et al., 2003], стремятся минимизировать ожидаемую стоимость, т.е. взвешенную по потерям вероятность неправильной классификации
на всех имеющихся и будущих данных. Эти методы наблюдают за расположением тестовых точек, и их целевые функции становятся монотонными в предсказательных энтропиях в тестовых точках. В [Kapoor et al., 2007] также включен эмпирический член ошибки

	MCMC	EP ($\approx$)	Лаплас ($\approx$)
MC	0	7.51 ± 2.51	41.57 ± 4.02
$\approx^2$	0.16 ± 0.05	7.43 ± 2.40	40.45 ± 3.67

Рисунок 2: Процентная ошибка аппроксимации (1 s.d.) для различных методов приближенного вывода (*столбцы*) и методов аппроксимации для оценки уравнения (4) (*строки*). Результаты показывают, что $\approx^2$ является очень точным приближением; EP приводит к некоторым потерям, а Лаплас - к значительно большим, что согласуется со сравнением, представленным в [Kuss and Rasmussen, 2005]. В наших экспериментах мы используем EP.

что может привести к патологическому поведению (мы исследуем это экспериментально). Эти подходы требуют больших вычислительных затрат, требуя (N $N_{\mathbf{x}y}$) обновления апостериорных данных. Кроме того, они должны знать местоположение тестовых данных (и, таким образом, являются трансдуктивными подходами); разработка индуктивного, теоретического алгоритма принятия решений является открытой и трудной проблемой, поскольку потребует дорогостоящего интегрирования по возможным распределениям тестовых данных.

Невероятностные Некоторые невероятностные методы имеют близкие аналоги информационно-теоретического активного обучения. Возможно, наиболее распространенным является активное обучение для SVM [Tong and Koller, 2001, Seung et al., 1992], где объем пространства версий (VS) используется в качестве косвенного показателя апостериорной энтропии. Если используется равномерное (неправильное) предшествование с детерминированной вероятностью классификации, то логарифмический объем VS и байесовская апостериорная энтропия фактически эквивалентны. Подобно тому, как байесовские апостериоры становятся неразрешимыми после наблюдения большого количества точек данных, VS может стать сложным. В работе [Tong and Koller, 2001] предложены методы аппроксимации VS с помощью простых фигур, таких как гиперсферы (их простейшая аппроксимация сводится к маргинальной выборке). Это очень похоже на аппроксимацию байесовской апостериорной оценки с помощью гауссовского распределения через аппроксимацию Лапласа или EP. В работе [Seung et al., 1992] эта проблема обойдена стороной, поскольку она работает с предсказаниями. В алгоритме Query by Committee (QBC) производится выборка параметров у VS (членов комитета), которые голосуют за результат каждого возможного $\boldsymbol{x}$. Выбирается $\boldsymbol{x}$ с наиболее сбалансированным голосованием; это называется "принципом максимального несогласия". Если BALD используется с выборочным апостериором, то запрос комитетом реализуется, но с вероятностной мерой разногласий. Детерминированный критерий голосования в QBC отбрасывает доверие к предсказаниям и поэтому может проявлять те же патологии, что и MES.

5 Эксперименты

Количественная оценка потерь при аппроксимации: Чтобы получить (5), мы сделали два приближения: выполнили приближенное умозаключение (^{1}) и аппроксимировали бинарную энтропию гауссовского CDF квадратичной экспонентой (^{2}). Оба этих метода могут быть заменены выборкой Монте-Карло, что позволит нам вычислить

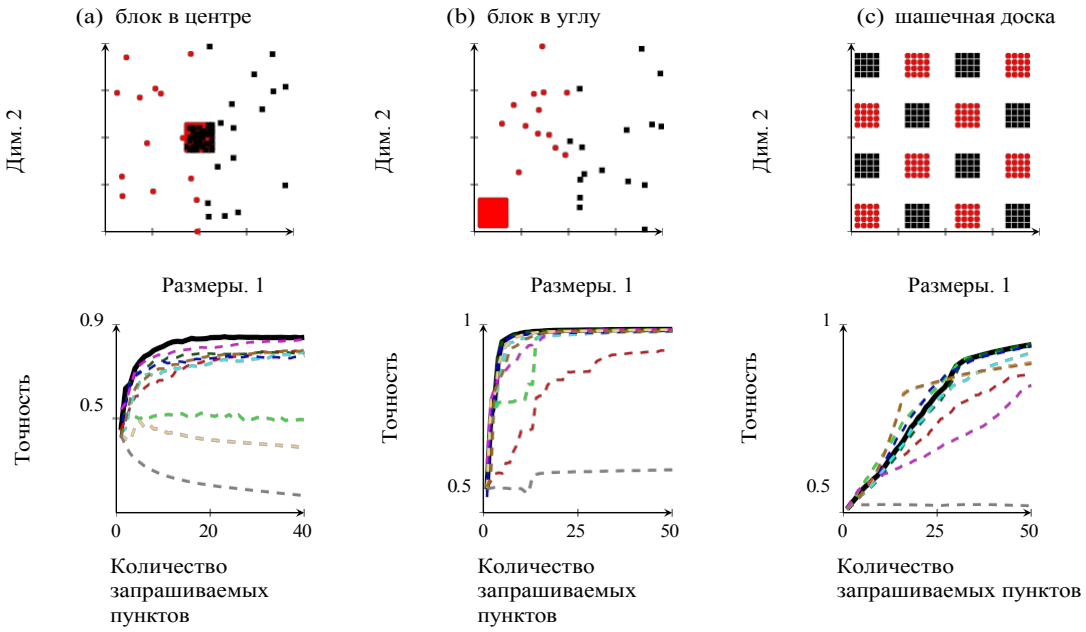

Рисунок 3: *Вверху:* оценка на искусственных наборах данных. Примеры двух классов показаны черными квадратами ( ) и красными кругами ( ). *Внизу:* Результаты активного обучения девятью методами: случайный запрос ( ), BALD( ), MES ( ), QBC с критерием голосования с 2 ( ) и 100 ( ) членами комитета, активная SVM ( ), IVM ( ), теоретические решения: [Kapoor et al., 2007] ( ), [Zhu et al., 2003] ( ) и эмпирическая ошибка ( ).

асимптотически несмещенная оценка ожидаемого информационного выигрыша. Используя обширное Монте-Карло в качестве "золотого стандарта", мы можем оценить, как много мы теряем, применяя эти приближения. Мы оцениваем ошибку аппроксимации как:

$$\frac{\max_{\boldsymbol{x}\in P} I(\boldsymbol{x}) - I(\arg\max_{\boldsymbol{x}\in P} \hat{I}(\boldsymbol{x}))}{\max_{\boldsymbol{x}\in P} I(\boldsymbol{x})} 100\% \qquad (8)$$

где I - цель, вычисленная методом Монте-Карло, $\hat{I}$ - приближенная цель. Использовался набор данных UCI *по раку*, результаты и обсуждение приведены на рис. 2.

Активное обучение на основе пула: Мы тестируем BALD для GPC и обучения предпочтениям в условиях пула, т.е. при выборе значений $\boldsymbol{x}$ из фиксированного набора точек данных. Хотя BALD может обобщить выбор непрерывных $\boldsymbol{x}$, это позволяет нам сравнивать с алгоритмами, которые не могут этого сделать. Мы сравниваем с восемью другими алгоритмами: случайной выборкой, MES, QBC (с 2 и 100 членами комитета), SVM с аппроксимацией пространства версий [Tong and Koller, 2001], теоретическими подходами в [Kapoor et al., 2007, Zhu et al., 2003] и прямой минимизацией

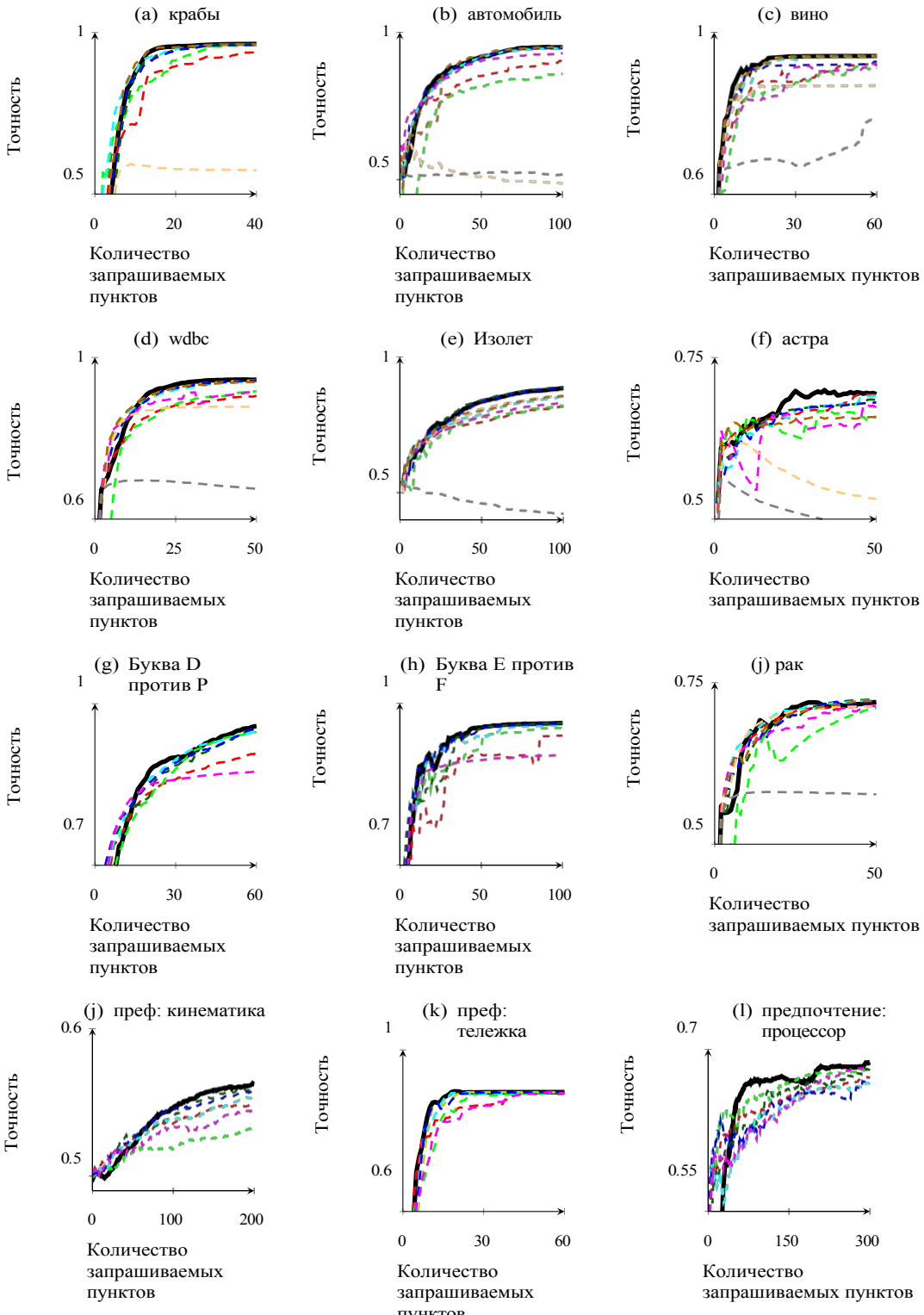
(a) крабы
(b) автомобиль
(c) вино
(d) wdbc
(e) Изолет
(f) астра
(g) Буква D против P
(h) Буква E против F
(j) рак
(j) преф: кинематика
(k) преф: тележка
(l) предпочтение: процессор
Точность
Количество запрашиваемых пунктов

Рисунок 4: Точность классификации тестового набора на наборах данных классификации и обучения предпочтениям. Использовались следующие методы: BALD( ), случайный запрос ( ), MES ( ), QBC с 2 (QBC_2 ,) и 100 (QBC_{100} ,) членами комитета, активная SVM ( ), IVM ( ), теоретический метод принятия решений [Kapoor et al., 2007] ( ), теоретический метод принятия решений [Zhu et al., 2003] ( ) и эмпирическая ошибка ( ). Теоретические методы требуют много времени для выполнения, поэтому они не были завершены для всех наборов данных. Графики (a-i) - наборы данных GPC, (j-l) - обучение предпочтениям.

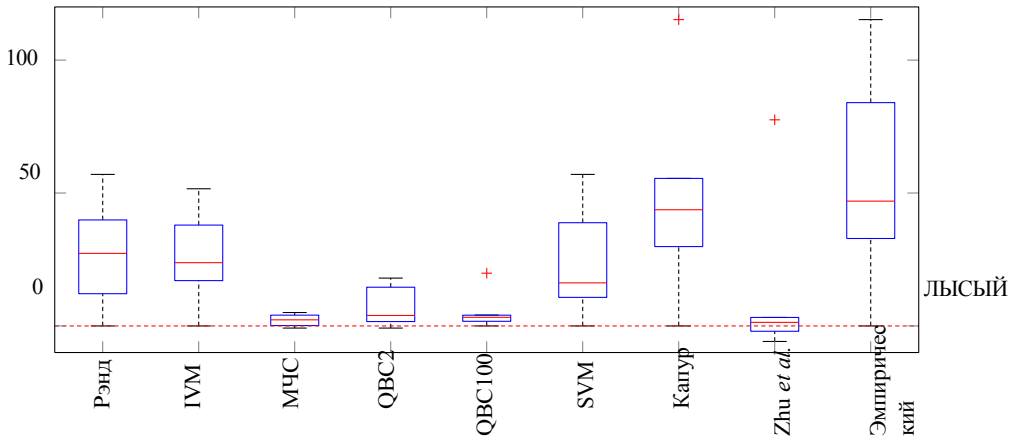

Рисунок 5: Сводка результатов всех экспериментов по классификации. *Ось y* обозначает количество дополнительных точек данных относительно BALD, необходимых для достижения по крайней мере 97,5% от предсказательной эффективности всего пула. Коробка" обозначает 25-75-й процентили, красная линия - медиану по набору данных, а "усы" - диапазон. Крестики обозначают выбросы *(> 2,7σ* от среднего значения). Положительные значения означают, что алгоритму требуется больше точек данных, чем BALD, для достижения той же производительности.

ожидаемая эмпирическая ошибка (последний метод не является широко используемым, но включен для анализа в [Kapoor et al., 2007]).

Мы рассматриваем три искусственных, но сложных набора данных. Первый из них, *блок в середине*, имеет блок зашумленных точек на границе принятия решения, второй, *блок в углу*, имеет блок неинформативных точек далеко от границы принятия решения: сильный алгоритм активного обучения должен избегать этих неинформативных областей. Третий набор данных похож на набор данных *"Шахматная доска"* в [Zhu et al., 2003] и предназначен для проверки возможностей алгоритма по поиску множества непересекающихся островков точек из одного класса. Три набора данных и результаты использования каждого алгоритма показаны на рис. 3.

Результаты также представлены на восьми классификационных датасетах UCI: *australia, crabs, vehicle, isolet, cancer, wine, wdbc* и *letter*. *Letter* - это мультиклассовый набор данных, для которого мы выбрали трудноразличимые буквы E против F и D против P. Для обучения предпочтениям мы используем наборы данных *cpu, cart* и *kinematics* regression. [1] обработанные для получения задачи предпочтения, как описано в [Chu and Ghahramani, 2005]. Результаты представлены на рис. 4, а на рис. 5 - их совокупность.

Обсуждение: На рис. 3 и 4 показано, что при использовании BALD мы добиваемся значительного превосходства над наивной случайной выборкой как в области классификации, так и в области обучения предпочтениям. По сравнению с другими алгоритмами активного обучения BALD постоянно является лучшим

[1]http://www.liacc.up.pt/ ltorgo/Regression/DataSets.html

лучший или один из лучших алгоритмов на всех наборах данных. На каждом отдельном наборе данных производительность BALD часто совпадает, поскольку мы сравниваем со многими методами, а более приближенные алгоритмы могут иметь хорошую производительность в разных условиях. На рис. 5 видно, что BALD обладает наилучшей общей производительностью; в среднем всем остальным методам требуется больше точек данных для достижения той же точности классификации. Наиболее близок к этому теоретико-решающий подход Чжу *и др.*: медианное увеличение количества требуемых точек данных составляет 1,4, а нулевое значение (т. е. эквивалентное BALD) находится в пределах интерквартильного интервала. Однако этот алгоритм требует гораздо больше вычислительного времени и имеет доступ к полному набору тестовых входов, чего нет у BALD. MES и QBC близки по производительности к BALD, но нулевая линия находится за пределами обоих интерквартильных диапазонов.

Как и ожидалось, MES плохо работает на зашумленном наборе данных (рис. 3(a)), поскольку он отбрасывает знание о шуме наблюдений. При нулевом шуме наблюдений он эквивалентен BALD, например, рис. 3(c). На многих реальных наборах данных MES работает так же хорошо, как и BALD, например, на рис. 4(b, e), что указывает на то, что эти наборы данных в основном не содержат шума.

IVM хорошо работает на рис. 3(c), но патологически - на 3(a); это связано с тем, что он смещает выборку в сторону точек только одного класса в зашумленном кластере, быстро, но искусственно уменьшая заднюю энтропию. Однако он также значительно хуже, чем BALD, работает на бесшумных (на что указывает высокая производительность MES) наборах данных, например, на рис. 4(b). Это означает, что апостериорная аппроксимация IVM или обновление ADF негативно сказываются на производительности алгоритма.

QBC часто дает лишь небольшое снижение производительности, приближение выборки часто не слишком вредно. Однако он плохо работает на зашумленном искусственном наборе данных (рис. 3(a)), поскольку критерий голосования не поддерживает понятие присущей неопределенности, как MES. Подход на основе SVM демонстрирует переменную производительность (он хорошо работает на рис. 4(d), но очень плохо - на 4(f)). На производительность сильно влияет используемая аппроксимация, для единообразия мы приводим здесь ту, которая дала наиболее стабильно хорошие результаты.

Подходы, основанные на теории принятия решений, иногда показывают хорошие результаты: в примере 3(c) они выбирают первые 16 точек из центра каждого кластера, поскольку на них влияют окружающие немаркированные точки. BALD не наблюдает за немаркированными точками, поэтому может не выбрать точки из центров. На рис. 5 видно, что BALD работает так же хорошо, как метод из [Zhu et al., 2003], и превосходит подход из [Kapoor et al., 2007], несмотря на отсутствие доступа к местоположению тестовых точек и значительно меньшие вычислительные затраты. Цель в [Kapoor et al., 2007] может быть неудачной, потому что один из членов в их объективной функции - эмпирическая ошибка. Вес, придаваемый этому члену, определяется относительными размерами обучающего и тестового множеств (и связанными с ними потерями). Прямая минимизация эмпирической ошибки обычно работает очень патологично, выбирая только "безопасные" точки. Когда метод в [Kapoor et al., 2007] присваивает этому члену слишком большой вес, он также может потерпеть неудачу.

Наконец, отметим, что BALD иногда плохо работает на первых нескольких точках данных (например, рис. 4(l)). Это может быть связано с тем, что гиперпараметры фиксированы на протяжении всех экспериментов, чтобы обеспечить справедливое сравнение с алгоритмами

не способна к обучению гиперпараметров. Это может означать, что при малом количестве данных GP-модель будет слишком хорошо подходить, что приведет к тому, что BALD будет выбирать ненормальные места для запросов. Поддержание распределения по гиперпараметрам может быть выполнено с помощью MCMC, но это значительно увеличивает время вычислений. Разработка общего метода, позволяющего делать это эффективно, является предметом дальнейшей работы. На практике обычно достаточно простой эвристики, например, случайного выбора нескольких первых точек и оптимизации гиперпараметров.

6 Выводы

Мы продемонстрировали метод, применяющий полный информационно-теоретический критерий активного обучения к классификации ГП, который, насколько известно авторам, делает наименьшее количество приближений на сегодняшний день и имеет столь же хорошую вычислительную сложность. Мы расширяем модель GPC для разработки нового ядра обучения предпочтениям, что позволяет нам применять наш алгоритм активного обучения непосредственно и в этой области. Метод может естественным образом обрабатывать активное обучение гиперпараметров ядра, что является сложной, в основном нерешенной проблемой, например, в активном обучении SVM. Примечательной особенностью нашего подхода является то, что он не зависит от используемых методов приближенного вывода. Это позволяет нам выбирать из целого ряда методов приближенного вывода, включая EP, аппроксимацию Лапласа, ADF или даже разреженное онлайн-обучение, и таким образом находить компромисс между вычислительной сложностью и точностью. Наши экспериментальные результаты выгодно отличаются от многих других методов активного обучения классификации и даже от методов теории принятия решений, которые имеют доступ к тестовым данным и требуют гораздо больше времени для вычислений.

Ссылки

[Bernardo, 1979] Бернардо, Дж. (1979). Ожидаемая информация как ожидаемая полезность. *Анналы статистики*, 7(3):686-690.

[Chu and Ghahramani, 2005] Chu, W. and Ghahramani, Z. (2005). Обучение предпочтениям с помощью гауссовских процессов. В *ICML*, страницы 137-144. ACM.

[Cover et al., 1991] Cover, T., Thomas, J., and Wiley, J. (1991). *Элементы теории информации*, том 6. Wiley Online Library.

[Dasgupta, 2005] Dasgupta, S. (2005). Анализ жадной стратегии активного обучения. In *NIPS*.

[Ertin et al.,] Ertin, E., Fisher, J., and Potter, L. Maximum mutual information principle for dynamic sensor query problems. В *обработке информации в сенсорных сетях*, Lecture Notes in Computer Science.

[Fuhrmann, 2003] Fuhrmann, D. (2003). *Активное тестирование систем наблюдения, или Игра в двадцать вопросов с радаром*. Центр оборонной технической информации.

[Fu¨rnkranz and Hu¨llermeier, 2003] Fu¨rnkranz, J. and Hu¨llermeier, E. (2003). Обучение парным предпочтениям и ранжирование. *Машинное обучение: ECML 2003*, страницы 145-156.

[Головин и Краузе, 2010] Головин, Д. и Краузе, А. (2010). Адаптивная субмодульность: Новый подход к активному обучению и стохастической оптимизации. In *COLT*.

[Guestrin et al., 2005] Guestrin, C., Krause, A., and Singh, A. P. (2005). Близкое к оптимальному размещение датчиков в гауссовских процессах. В *материалах 22-й международной конференции по машинному обучению*, ICML '05, стр. 265-272, Нью-Йорк, штат Нью-Йорк, США. ACM.

[Heckerman et al., 1995] Heckerman, D., Breese, J., and Rommelse, K. (1995). Устранение неполадок в условиях неопределенности. *Communications of the ACM*, 38(3):27-41.

[Huang et al., 2010] Huang, S., Jin, R., and Zhou, Z. (2010). Активное обучение путем запроса информативных и репрезентативных примеров. *Advances in neural information processing systems*, 23:892-900.

[Kapoor et al., 2007] Kapoor, A., Horvitz, E., and Basu, S. (2007). Selective su- pervision: Направление контролируемого обучения с помощью теоретико-решающего активного обучения. In *IJCAI*.

[Krause et al., 2006] Krause, A., Guestrin, C., Gupta, A., and Kleinberg, J. (2006). Околооптимальное размещение датчиков: Максимизация информации при минимизации затрат на связь. В *материалах 5-й международной конференции по обработке информации в сенсорных сетях*, стр. 2-10. ACM.

[Krishnapuram et al.,] Krishnapuram, B., Williams, D., Xue, Y., Hartemink, A., Carin, L., and Figueiredo, M. On semi-supervised classification. *NIPS*.

[Kuss and Rasmussen, 2005] Kuss, M. and Rasmussen, C. E. (2005). Оценка аппроксимаций для классификации гауссовских процессов. In *NIPS*. MIT Press.

[Lawrence et al., 2003] Lawrence, N., Seeger, M., and Herbrich, R. (2003). Быстрые разреженные методы гауссовых процессов: Информативная векторная машина. *Advances in neural information processing systems*, pages 625-632.

[Lindley, 1956] Линдли, Д. (1956). О мере информации, предоставляемой экспериментом. *Анналы математической статистики*, 27(4):986-1005.

[MacKay, 1992] MacKay, D. (1992). Основанные на информации целевые функции для активного выбора данных. *Нейронные вычисления*, 4(4):590-604.

[Panzeri and Petersen, 2007] Panzeri, S., S. R. M. M. and Petersen, R. (2007). Коррекция проблемы смещения выборки в информационных мерах спайк-трейна. *Журнал нейрофизиологии*, 98(3):1064.

[Rasmussen and Williams, 2005] Rasmussen, C. and Williams, C. (2005). *Gaus- sian Processes for Machine Learning*. The MIT Press.

[Roy and McCallum, 2001] Roy, N. and McCallum, A. (2001). На пути к оптимальному активному обучению через выборочную оценку уменьшения ошибок. В *ICML*, страницы 441-448.

[Sebastiani and Wynn, 2000] Sebastiani, P. and Wynn, H. (2000). Выборка с максимальной энтропией и оптимальный байесовский экспериментальный дизайн. *Журнал Королевского статистического общества: Серия B (Статистическая методология)*, 62(1):145-157.

[Seung et al., 1992] Seung, H., Opper, M., and Sompolinsky, H. (1992). Запрос по комитету. В *COLT*, страницы 287-294. ACM.

[Tong and Koller, 2001] Tong, S. and Koller, D. (2001). Активное обучение машины векторов поддержки с приложениями к классификации текстов. *Журнал исследований машинного обучения*, 2:45-66.

[Zhu et al., 2003] Zhu, X., Ghahramani, Z., and Lafferty, J. (2003). Комбинирование активного обучения и полунаблюдаемого обучения с использованием гауссовых полей и хар-монических функций. *Семинар ICML 2003 "Континуум от меченых к немеченым данным в машинном обучении и добыче данных*".

ПРИЛОЖЕНИЕ - ДОПОЛНИТЕЛЬНЫЕ МАТЕРИАЛЫ

Разложение Тейлора для аппроксимации $^2 \approx$

Мы выполняем разложение Тейлора для ln H[Φ(*x*)] следующим образом:

$$f(x) = f(0) + \frac{f'(0)x}{1!} + \frac{f''(0)x^2}{2!} + \dots$$

$$f(x) = \ln \mathrm{H}[\Phi(x)]$$

$$f'(x) = \frac{1}{\ln 2}\frac{\Phi'(x)}{\mathrm{H}[\Phi(x)]}[\ln\Phi(x) - \ln(1-\Phi(x))]$$

$$f''(x) = \frac{1}{\ln 2}\frac{\Phi'(x)^2}{\mathrm{H}[\Phi(x)]^2}[\ln\Phi(x) - \ln(1-\Phi(x))] + \frac{1}{\ln 2}\frac{\Phi''(x)}{\mathrm{H}[\Phi(x)]}[\ln\Phi(x) - \ln(1-\Phi(x))] - \frac{1}{\ln 2}\frac{\Phi'(x)^2}{\mathrm{H}[\Phi(x)]}\left(\frac{1}{\Phi(x)} + \frac{1}{(1-\Phi(x))}\right)$$

$$\therefore \ln \mathrm{H}[\Phi(x)] = 1 - \frac{1}{\pi \ln 2}x^2 + O(x)^4$$

Поскольку функция четная, мы можем убедиться, что член x^3 будет равен нулю. Таким образом, экспоненцируя, мы получаем приближение до $O(x^4)$:

$$\mathrm{H}[\Phi(x)]^2 \approx \exp\left(-\frac{x^2}{\pi \ln 2}\right)$$

Ядро предпочтений

Среднее μ_{pref}, и ковариационная функция k_{pref} ГП над g могут быть вычислены из среднего и ковариации $f \sim$ ГП(μ, k) следующим образом:

$$\begin{aligned} k_{\text{pref}}([\boldsymbol{u}_i, \boldsymbol{v}_i], [\boldsymbol{u}_j, \boldsymbol{v}_j]) &= Cov[g(\boldsymbol{u}_i, \boldsymbol{v}_i), g(\boldsymbol{u}_j, \boldsymbol{v}_j)]. \\ &= Cov[(f(\boldsymbol{u}_i) - f(\boldsymbol{v}_i)), (f(\boldsymbol{u}_i) - f(\boldsymbol{v}_i))]. \\ &= \mathrm{E}[(f(\boldsymbol{u}_i) - f(\boldsymbol{v}_i)) - (f(\boldsymbol{u}_i) - f(\boldsymbol{v}_i))] \\ &\quad - (\mu(\boldsymbol{u}_i) - \mu(\boldsymbol{v}_i))(\mu(\boldsymbol{v}_j) - \mu(\boldsymbol{u}_i)) \\ &= k(\boldsymbol{u}_i, \boldsymbol{u}_j) + k(\boldsymbol{v}_i, \boldsymbol{v})_j \\ &\quad - k(\boldsymbol{u}_i, \boldsymbol{v}_j) - k(\boldsymbol{v}_i, \boldsymbol{u})_j \end{aligned} \tag{9}$$

$$\begin{aligned} \mu_{\text{pref}}([\boldsymbol{u}, \boldsymbol{v}]) &= \mathrm{E}[g([\boldsymbol{u}, \boldsymbol{v}])] = \mathrm{E}[f(\boldsymbol{u}) - f(\boldsymbol{v})] \\ &= \mu(\boldsymbol{u}) - \mu(\boldsymbol{v}) \end{aligned} \tag{10}$$