Семинар: вероятностная постановка задачи в машинном обучении. Логистическая регрессия

Табаченков Андрей Михайлович
Федоров Артем Максимович Афанасьев Глеб Ильич

Октябрь, 2025

1 Введение

Данный семинар предназначен познакомить слушателей с вероятностной парадигмой машинного обучения. Для начала введем условия решаемой задачи в общем виде:

1. Дано некоторое множество Ω объектов ω.
2. Каждый объект ω изучаемого множества может быть представлен в виде набора из s численных признаков $g_i(\omega)$, где $g_i : \Omega \to \mathbb{R}, i = \overline{1,s}$.
3. Каждый объект ω также обладает некоторыми скрытыми свойствами k, не представленными в признаковом описании $\vec{g}(\omega) = (g_1(\omega), ..., g_s(\omega))$.
4. В задаче также определена выборка $(X_N, K_N) = \{(x_1, k_1), ..., (x_N, k_N)\} \subset \vec{g}(\Omega) \times K$, по которой требуется минимизировать риск $\mathcal{R}(q) \to \min\limits_{q \in M}$, M - класс стратегий.

Таким образом, целью задачи является восстановить закон $q : \vec{g} \to K$, по которому из имеющегося признакового описания $x = g(\omega)$ каждого объекта мы определяем сокрытую информацию о нем.

Пример 1
Задача: предсказать потребность в ремонте станка

- Ω – все возможные состояния станка
- g_i – данные снятые с датчика
- $K = \{0, 1\}$ – сломается ли станок в ближайшую неделю

Пример 2
Задача: определить ущерб после стихийного бедствия

- Ω – стихийное бедствие
- g_i – его параметры (место, время, оценка силы ...)
- $K = [0, +\infty]$ – стоимость ущерба инфраструктуре и частной собственности

Пример 3
Задача: по EEG и MEG человека определить приказ для робота-манипулятора

- Ω – все возможные мысли человека
- g_i – получаемые временные ряды с каждого датчика
- $K = Action space$ – пространство действий (схватить, повернуть влево, сдвинуться)

2 Вероятностная парадигма машинного обучения

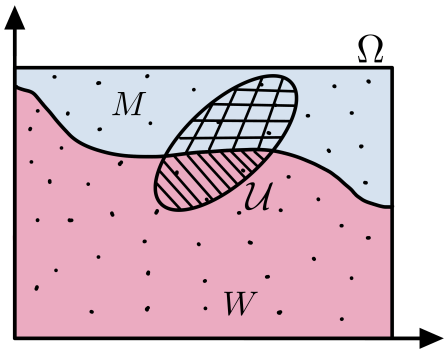

Рис. 1: Прообразы всех возможных наблюдений эксперимента. ($\cdot$ – объект из X_N, M - man, W - woman, $\mathcal{U} - g^{-1}(x)$)

Рассмотрим задачу определения пола человека по его фотографии. Для этого будет подразумеваться, что была определена случайная выборка людей (например с улицы) и фотосъемка проводилась в разных условиях (погода, разные камеры, разная одежда на людях) - каждый рецензент проходил ровно один раз. Тогда $\Omega = \Omega_1 \times \Omega_2$ - все возможные обстоятельства, повлиявшие на пол человека и его внешний вид, а также все условия, влияющие на фотографию человека. $K = \{M, W\}$ – есть пол человека.

Пусть у нас есть наблюдение $x \in \vec{g}(\Omega)$. Рассмотрим прообраз $\mathcal{U} = \vec{g}^{-1}(x) \subset \Omega$. Обозначим за Ω_M и Ω_W - подмножества Ω с фотографиями мужчин и фотографиями женщин. Для простоты можно предположить, что $\Omega_M \cap \Omega_W = \varnothing, \Omega_M \cup \Omega_W = \Omega$.

Заметим, что вообще говоря, пересечения $\mathcal{U} \cap \Omega_M$ и $\mathcal{U} \cap \Omega_W$ оба могут быть непусты 1, так как отображение $\vec{g}$ не обязательно биективно. Значит, мы, вообще говоря, не можем построить в общем случае полностью правильный классификатор пола для данной задачи, мы лишь можем только рассматривать вероятности $\mathbb{P}(M|\mathcal{U}) = \frac{|\mathcal{U} \cap \Omega_M|}{|\mathcal{U}|}, \mathbb{P}(W|\mathcal{U}) = \frac{|\mathcal{U} \cap \Omega_W|}{|\mathcal{U}|}$ и предполагать, что $k = M \Leftrightarrow \mathbb{P}(M|\mathcal{U}) \gg \mathbb{P}(W|\mathcal{U})$. Таким образом, в рамках вероятностной парадигмы мы стремимся устранить этот недостаток, оценивая соответствующие условные распределения.

Для соответствия задачи машинного обучения вероятностной парадигме достаточно потребовать:

1. Измеримость функций g_i.
2. Наличие в выборке (X_N, K_N) всех объектов из генеральной совокупности исследуемого множества.
3. Одновременное существование вероятностей
$$\mathbb{P}(x,k), \mathbb{P}(x), \mathbb{P}(k), \mathbb{P}(x|k), \mathbb{P}(k|x)$$
4. Существование функции потерь $\mathcal{L} : K \times X \times K \to \mathbb{R}$, где $\mathcal{L}$ - определена и интегрируема по Лебегу.
5. Представимость минимизируемого риска в следующем виде: $\mathcal{R}(q) = \mathbb{E}_{X,K}[\mathcal{L}(q(x), x, k)] \to \min\limits_{q \in M}$.

$$\mathbb{E}_{X,K}\left[\mathcal{L}(q(x), k, x)\right] = \int\limits_X \int\limits_K p(x,k)\mathcal{L}(q(x), x, k)dxdk =$$
$$= \int\limits_X p(x)dx \int\limits_K p(k \mid x)\mathcal{L}(q(x), x, k)dk$$

Данная задача называется генеративной задачей машинного обучения, лежащей в основе Байесовского метода распознавания и используемая всюду, от *Sklearn* до глубокого обучения. Основная причина, почему вероятностная модель (говоря простыми словами, модель оценивающая меры множеств) настолько удобна заключается в том, что мы способны автоматически обрабатывать неточности данных, задавать аналитически плоскость лосса и прочие приятные фишки аппарата теории вероятностей. Огромным преимуществом генеративных моделей также является то, что мы обучаем совместное распределение целевой метки и наблюдений, следовательно мы можем не только решать задачу нахождения k, но и генерировать объекты из X.

3 Проблемы генеративных моделей

Одной из основных проблем генеративных моделей является обязательное условие существования априорных распределений, хотя моделирование таких априорных распределений не всегда возможно и зачастую трудозатратно. Кроме того, если говорить отдельно о $\mathbb{P}(x)$, то зачастую от нас даже не требуется понимать вероятностную природу множества X. Далее приведем примеры, иллюстрирующие проблемы генеративных моделей, при использовании их на практике:

1. **Пример Не Байесовской задачи:** Пусть вы работаете на военной базе главным аналитиком. Вам поручено разработать систему "Свой-чужой основанную на визуальных и ЭРЛС данных о самолетах противника и союзников. Имея опыт работы с данными вы быстро понимаете, что речь идет о вероятностной задаче бинарной классификации. Однако можно ли применить в данном случае генеративный подход?

   Оказывается, что нет! Первым же делом возникает вопрос: что такое априорная вероятность события "Вражеский самолет"? Как мы можем оценить вероятность события, что в действительности не может происходить в мирное время, а если и произошло, то это явно признак "неординарной"ситуации. Более формально ситуацию может разъяснить теория вероятностей: мы считаем вероятность события равной $p \iff$ при бесконечном проведении независимых опытов, событие будет воспроизводиться с частотой p. Соответственно, в данной задаче такой вероятности быть просто не может в мирное время, в котором система действовать и обязана.

2. **Пример неслучайной выборки:** Однако не всегда проблема кроется исключительно в задаче. Пусть задача состоит в оценке состояния оборудования завода по численным признакам с датчиков. Из огромного числа логов заказчик выбирает 100 наиболее характерных примеров состояний и таргет значений. Тогда такая ситуация так же становится не разрешимой с точки зрения генеративных моделей, так как неслучайная выборка не позволяет нам оценить априорные распределения событий.

Поэтому, чтобы остаться в рамках вероятностной парадигмы машинного обучения, принято переходить к рассмотрению дискриминативной задачи $\mathbb{E}_K[\mathcal{L}(q(x), x, k)|X] \to \min\limits_{q \in M}$, то есть мы стремимся по наблюдению x предсказать только k, забывая о совместных распределениях и вероятностной природе X, что также требует меньше данных и позволяет бороться с проблемами, описанными выше.

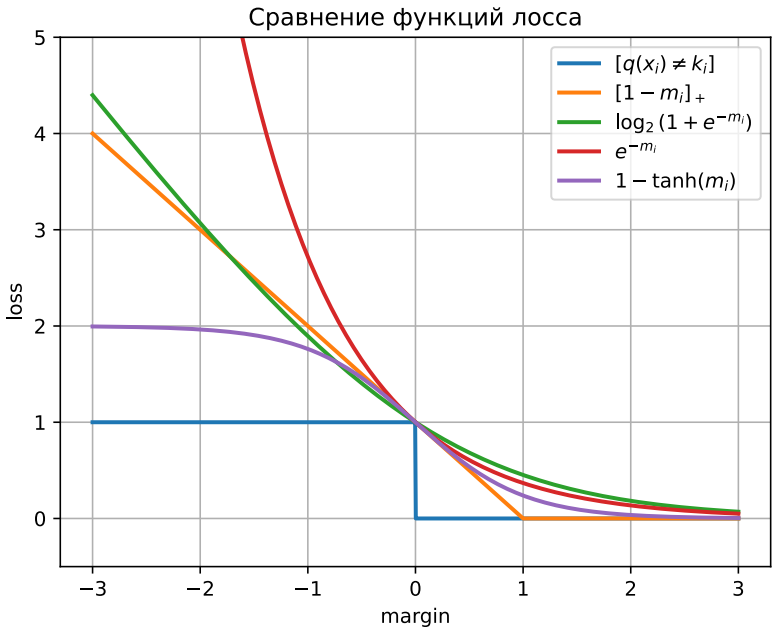

Рис. 2: Примеры лосс функций

4 Задача бинарной классификации

Пусть теперь множество таргетов $K = \{-1, 1\}$, то есть мы рассматриваем задачу бинарной классификации. Возникает вопрос, какую функцию потерь следует использовать. Наиболее простой способ - использовать accuracy:

$$\mathcal{R}(q) = \frac{1}{N} \sum_{i=1}^{N} [q(x_i) \neq k_i]$$

Несмотря на простоту и интуитивную понятность, есть ряд недостатков.

1. Такой лосс не является дифференцируемым в каноническом смысле $\Rightarrow$ оптимизация не более чем методами 0-порядка.

2. Поверхность данного лосса очень сложная с "постоянными"локальными минимумами (см. рис. 3).

Введем в рассмотрение отступы $m_i = q(x_i)k_i$ (в данном случае под $q(\cdot) : X \to \mathbb{R}$ также понимается дискриминативная функция). Тогда мы можем приблизить сверху ранее описанную функцию потерь другой функций

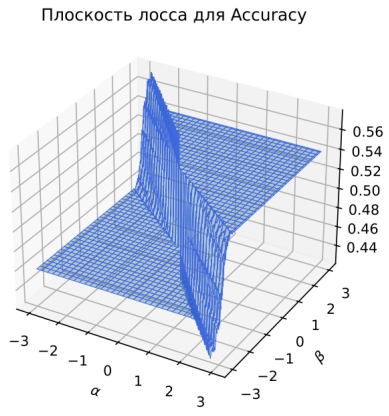

Рис. 3: Демонстрация accuracy loss функции в пространстве параметров

потерь (рис. 2), которая уже не обладает вышеописанными проблемами:

$$\mathcal{R}(q) \leq \hat{\mathcal{R}}(q) = \frac{1}{N}\sum_{i=1}^{N}\hat{\mathcal{L}}(m_i)$$

Однако, далеко не факт, что мы можем использовать данную функцию потерь в рамках вероятностной парадигмы.

5 Критерий вероятностной природы лосса

Чтобы понять, какими свойствами должна обладать функция потерь в вероятностной парадигме в задаче бинарной классификации, рассмотрим для начала следующий пример.

Пусть поставлена задача рекомендации пользователю рекламы. Аналитики компании уже проанализировали систему Ω и построена функция $g : \Omega \to \{1, 2, 3, \dots l\}$, что показывает некоторый категориальный признак для пары "реклама $\times$ пользователь". По полученной выборке (X_N, K_N) обучить стратегию определять вероятность того, что пользователь перейдет по рекламе. Тогда очевидно:

$$\sum_{i=1}^{N}\frac{\mathbf{1}\{Click|x_i = j\})}{\#(x = j)} = \frac{\text{кликов при } j}{\text{всего наблюдений } j} \xrightarrow[n\to\infty]{} p(Click \,|\, j)$$

Будем считать, что обучающая выборка нашей модели настолько велика, насколько мы захотим $(N \to \infty)$. Тогда для любого наблюдаемого значения $j \in \{1, 2, 3, \dots l\}$ мы сможем получить вероятность $p(\text{пользователь кликнул} \,|\, j) =$

$1 - p(\text{пользователь не кликнул} \mid j)$ и соответственно будем хотеть от нашей вероятностной модели выдавать ровно такие же значения на наблюдения x:

$$q(x = j) = p(Click|j)$$

Вернемся теперь к общему случаю. Как итог, нам хотелось бы, чтобы оптимальная модель $q \in M$ приближала вероятность положительного класса для каждого из наблюдений (при достаточно большом N).

$$\min_{q \in M} \hat{\mathcal{R}}(q) = \min_{q \in M} \mathbb{E}[\hat{\mathcal{L}}(q(x), x, k)|x] = \min_{q \in M} \mathbb{E}[m_i|x]$$

$$\arg\min_{q \in M} \hat{\mathcal{R}}(q) \xrightarrow[N \to +\infty]{} \mathbb{P}(k = 1|\cdot)$$

5.1 Интерпретация данного результата на более сложных моделях

Пусть для пространства признаков выполнено условие "континуальности"(непрерывности из лекций Воронцова). Тогда мы фактически говорим, что отображение g ставит близкие объекты относительно изначального Ω близко друг другу и в пространстве признаков. Тогда мы можем говорить, что при его разбитии на малые области (**квантизация** X), разумная модель для каждого объекта будет выдавать вероятности, равные отношению классов в данной области.

$$\rho(x, c) \leq \delta \Rightarrow \mathcal{R} = \sum \tilde{\mathcal{L}}(x) \approx \sum \left[\tilde{\mathcal{L}}(c) + \bar{\mathcal{O}}(x - c)\right] \approx \sum \tilde{\mathcal{L}}(c)$$

Функция потерь на близких объектах будет практически равна одному и тому же, а значит и при обучении на таких объектах должны в «идеале» выдаваться почти одинаковые вероятности. Но тогда

$$\text{доля классов} = \mathbb{P}(k = 1|\rho(x, c) \leq \delta) = \int\limits_{\rho(x,c) \leq \delta} p(k = 1|x) d\mathbb{P}(x|\rho(x, c) \leq \delta) \approx$$

$$\approx p(k = 1|c) \int\limits_{\rho(x,c) \leq \delta} d\mathbb{P}(x|\rho(x, c) \leq \delta) = p(k = 1|c) \Rightarrow$$

$$\Rightarrow p(k = 1|x) \approx p(k = 1|c) = \text{доля классов} \square$$

5.2 Примеры проверки по критерию

Задача 1: Задача бинарной классификации: $\mathcal{L}(q,x,k) = (q(x) - [k=+1 \,|\, x])^2$ – MSE.

$$arg\min_q \mathbb{E}(\mathcal{L}(q,x,k) \,|\, x) = arg\min_q \left[(q(x)-1)^2 \underbrace{p(k=+1|x)}_{p} + q^2(x) \underbrace{p(k=-1|x)}_{1-p} \right]$$

$$\frac{\partial}{\partial q(x)} \mathbb{E}(\mathcal{L}(q,x,k) \,|\, x) = 2(q(x)-1)p + 2q(x)(1-p) = 0 \Rightarrow q(x) = p(k=+1 \,|\, x)$$

Задача 2: Задача бинарной классификации: $\mathcal{L}(q,x,k) = |q(x) - [k=+1 \,|\, x]|$ – MAE.

$$arg\min_q \mathbb{E}(\mathcal{L}(q,x,k) \,|\, x) = arg\min_q \left[|q(x)-1| \underbrace{p(k=+1|x)}_{p} + |q(x)| \underbrace{p(k=-1|x)}_{1-p} \right]$$

$$arg\min_q [p(1-q(x)) + q(x)(1-p)] = \arg\min_q [p + q(x) - 2pq(x)] \Rightarrow$$

$$\Rightarrow q(x) \neq p(k=+1 \,|\, x)$$

6 Задание функции потерь из вероятностного распределения

Теперь будем рассматривать только линейные классификаторы. Тогда в общем виде получаем, что $q(x|w) = f(\langle x, w \rangle)$, где функция f нормирует выход линейной функции на отрезок $[0,1]$. В качестве функции риска будем использовать оценку правдоподобия:

$$\mathcal{R} = \sum_{i=1}^{N} log\, p(k_i|x_i,w) \Leftrightarrow L = exp\{\mathcal{R}\} = \prod_{i=1}^{N} p(k_i|x_i,w)$$
$$p(k_i|x_i,w) := q(x_i|w)[k_i=+1] + (1-q(x_i|w))[k_i=-1]$$

Заметим, что использование оценки правдоподобия удовлетворяет критерию вероятностной природы лосса:

$$\arg\max_{q\in M} \mathbb{E}[\mathcal{L}(q,x,k)|x] = \arg\max_{q\in M} \mathbb{E}[log(q(x))[k=+1|x] + log(1-q(x))[k=-1|x]] =$$
$$= \arg\max_{q\in M} [log(q(x))p(k=+1|x) + log(1-q(x))(1-p(k=+1|x))]$$

Тогда

$$\frac{\partial}{\partial q(x)} \mathbb{E}(\mathcal{L}(q,x,k) \,|\, x) = \frac{p(k=+1|x)}{q(x)} - \frac{1-p(k=+1|x)}{1-q(x)} = 0 \Leftrightarrow q(x) = p(k=+1|x)$$

6.1 Логистическая регрессия

Делаем теперь еще одно предположение - пусть вероятность $p(k|x)$ есть Бернуллевское параметрическое семейство $Bern(p) = Bern(p(\langle x, w\rangle))$. Как будет доказано в конце семинара с использованием GLM (Generalised Linear Machine), функция f в таком случае должна иметь специфический вид, а именно $f(t) = \sigma(t) = \frac{1}{1+exp(-t)}$. Но тогда

$$-logL = -R = -\sum_{i=1}^{N}[log\ \sigma(\langle x_i, w\rangle)[k_i = +1] + log(1 - \sigma(\langle x_i, w\rangle))[k_i = -1]] =$$

$$= -\sum_{i=1}^{N}[log\ \sigma(\langle x_i, w\rangle)[k_i = +1] + log\ \sigma(-\langle x_i, w\rangle)[k_i = -1]] =$$

$$= -\sum_{i=1}^{N} log\ \sigma(k_i \langle x_i, w\rangle) = \sum_{i=1}^{N} log(1 + exp(k_i \langle x_i, w\rangle))$$

7 Оценка качества восстановления вероятностей модели

Построение вероятностной задачи не является панацеей, так как модель может быть недообучена или переобучена на задачу, выбранное распределение таргет переменных которой может отличаться от наших ожиданий, тем самым вся модель будет предсказывать совершенно неестественные распределения на объектах. Хотелось бы вывести (аналитически) более эмпирическую оценку качества вероятностной модели, нежели «суррогатный» эмпирический риск $\mathcal{R}$.

Рассмотрим такую ситуацию: для задачи бинарной классификации определим набор моделей, каждая обучаемая на различные гипотезы об апостериорном распределении таргета и с различными гиперпараметрами. Тогда для каждой модели строим "диаграммы калибровки":

1. Разбиваем ось $[0, 1]$ на бины измельчением $\{a_1, a_2, \dots a_\ell\}$, $[a_i, a_{i+1}] = \mathbb{B}_i$
2. Для каждого объекта высчитываем оценку вероятности $+1$ класса $p_i \in \mathbb{B}_i$
3. Для каждого бина считаем отношение попавших в него действительно $+1$ классов к общему числу попавших.

Такие диаграммы выполняют вполне осязаемую функцию: показать, насколько сильно предсказанная вероятность модели отличается от действительной выборочной.

По аналогии с примером про компактные выборки, нетрудно показать, что, так как распределения $+1$ классов у объектов в бине примерно одни

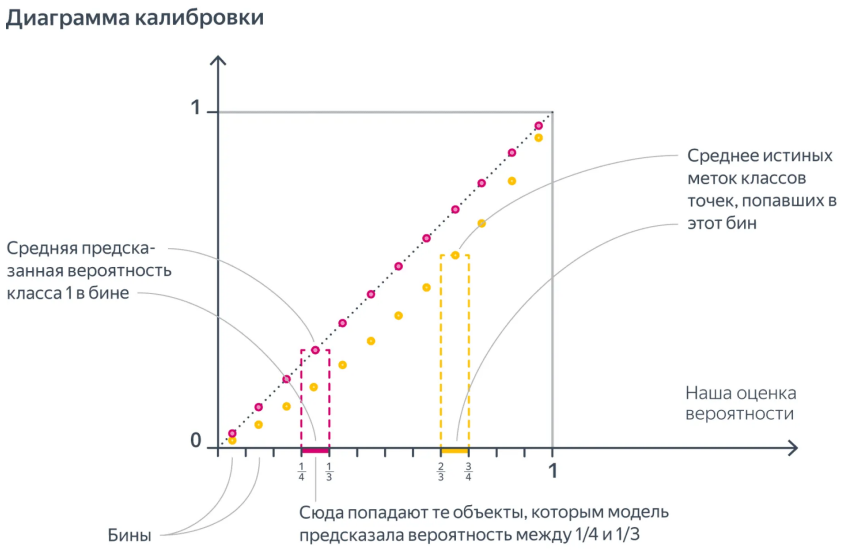

Рис. 4: Диаграмма калибровки

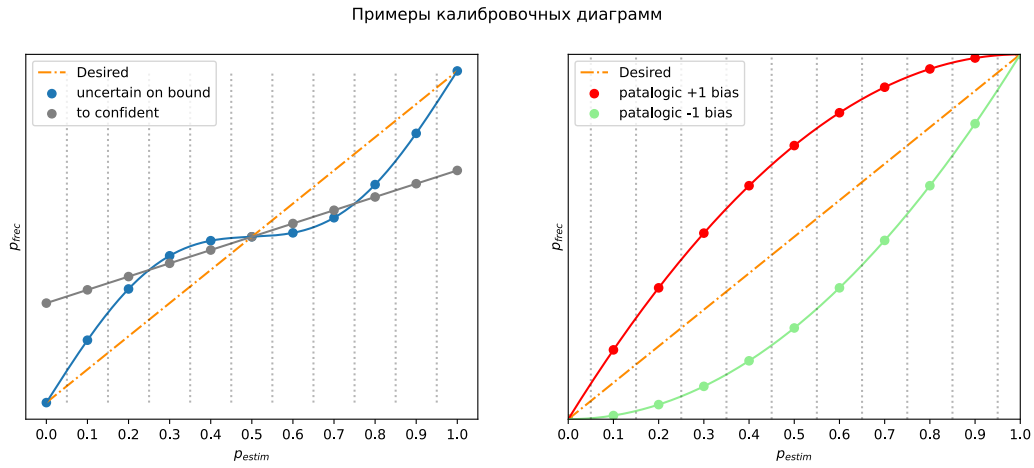

Рис. 5: Примеры калибровачных диаграмм

и те же, то тогда (если модель оценивает распределения практически идеально) они все должны быть примерно равны посчитанному отношению. Именно на этом и основан принцип таких диаграмм.
На примере (рис. 5) отчетливо видны 5 ситуаций:

1. **Desired:** Идеальное поведение классификатора, к которому мы стре-

мимся при обучении вероятностной модели

2. **Uncertain on bound:** модель, что отдает сильное предпочтение сильно отдаленным объектам. В таком случае крайне вероятно переобучение.

3. **To confident:** модель предсказывает каждый из классов слишком уверенно

4. **patalogic +1 bias:** модель старательно отдает предпочтение -1 классу. Эта ситуация является поводом задуматься о смене либо модели данных, либо о классе моделей мл

5. **patalogic -1 bias:** противоположная ситуация предыдущей.

7.1 Функционал качества оценки вероятности

В случае поведения вашей модели по сценарию **patalogic +1 bias** или **patalogic -1 bias** скорее всего вашему решению мало что поможет, кроме как смены архитектуры или рассматриваемой модели данных. Однако в остальных случаях есть шанс поднастроить модель, "просмотреть паттерн ее ошибок и вручную изменить ей выходные оценки". Семейство таких методов называется «калибровками». Для этого введем функционал качества калибровки **Expected/Maximum calibration error** по аналогии с диаграммой:

$$\mathbf{ECE} = \sum_{i=1}^{\ell} \frac{|\mathbb{B}_i|}{N} \left| \frac{1}{|\mathbb{B}_i|} \sum_{x \in \mathbb{B}_i} q(x) - \frac{1}{|\mathbb{B}_i|} \sum_{x \in \mathbb{B}_i} \mathbb{1}\{y_i = +1\} \right|$$

$$\mathbf{MCE} = \max_{i \in \overline{1,\ell}} \frac{|\mathbb{B}_i|}{N} \left| \frac{1}{|\mathbb{B}_i|} \sum_{x \in \mathbb{B}_i} q(x) - \frac{1}{|\mathbb{B}_i|} \sum_{x \in \mathbb{B}_i} \mathbb{1}\{y_i = +1\} \right|$$

Это лишь метрика качества, но не функционал для оптимизации, так как изменение модели влечет изменение и правой части, из-за чего его использование ограничивается оценкой качества.

7.2 Методы калибровки вероятностей

Методы калибровки предлагают не изменять модель напрямую. Напротив, они фиксируют веса модели, в то же время используя ее выход как новый признак объектов, что в теории позволяет усложнить модель и сделать ее более "нелинейной". Более того, в таком случае мы получаем возможность сделать вероятностную модель из не вероятностной. Всего существует 3 наиболее применимых и популярных подхода:

- **Калибровка Платта** – самая очевидная и широко применяемая, основанная на "логистической регрессии"поверх выхода классификатора.

- **Isotonic regression** – классическая задача восстановления функции распределения поверх стратегии q через изотоническую регрессию.
- **Гистограммная калибровка** – упрощенный вариант изотонической регрессии, с фиксированными границами $\mathbb{B}_i$

7.2.1 Калибровка Платта

Изначально разрабатываемая для получения вероятности из марджина невероятностных моделей (в частности **SVM**), представляет собой по сути применение еще одного линейного слоя (линейной модели) от одного аргумента с применением сигмоиды. Обучаемыми параметрами являются α и β – масштаб и сдвиг:

$$p(k_i = +1 \,|\, x_i) = \sigma\left(\alpha q(x_i) + \beta\right) = \frac{1}{1 + \exp\{-\alpha q(x_i) - \beta\}}$$

$$\Longrightarrow \sum_{i=1}^{N} \log\left(1 + \exp\{-\alpha q(x_i) - \beta\}\right) \longrightarrow \min_{\alpha,\beta \in \mathbb{R}}$$

Усложнение модели Платта: вместо сигмоиды можно использовать какую-нибудь другую функцию, дающую на выходе число от 0 до 1, например функцию плотности бета распределения (рис. 6):

$$p(q(x) \,|\, k = +1, \alpha_1, \beta_1) = \mathcal{B}(q(x), \alpha_1, \beta_1) = \frac{q(x)^{\alpha_1 - 1}\left(1 - q(x)\right)^{\beta_1 - 1}}{\mathcal{B}(\alpha_1, \beta_1)}$$

$$p(q(x) \,|\, k = -1, \alpha_2, \beta_2) = \mathcal{B}(q(x), \alpha_2, \beta_2) = \frac{q(x)^{\alpha_2 - 1}\left(1 - q(x)\right)^{\beta_2 - 1}}{\mathcal{B}(\alpha_2, \beta_2)}$$

Руководствуясь леммой Неймана-Пирсона, найдем отношение плотностей вероятности и обучим параметры на максимум правдоподобия:

$$LR = \frac{q(x)^{\alpha_1 - 1}\left(1 - q(x)\right)^{\beta_1 - 1}}{q(x)^{\alpha_2 - 1}\left(1 - q(x)\right)^{\beta_2 - 1}} \cdot \frac{\mathcal{B}(\alpha_2, \beta_2)}{\mathcal{B}(\alpha_1, \beta_1)} = \frac{q(x)^a}{\left(1 - q(x)\right)^b} K \simeq \frac{q(x)^a}{\left(1 - q(x)\right)^b} e^{-c}$$

Здесь мы приблизили отношение Бета функций экспонентой. Теперь же превратим отношение правдоподобия в вероятность, $a \geq 0,\ b \geq 0$, и получим классический вид **Бета калибровки**:

$$p(k = +1 \,|\, q(x), a, b, c) = \frac{1}{1 + LR^{-1}} = \frac{1}{1 + e^c \diagup \frac{q(x)^a}{(1 - q(x))^b}}$$

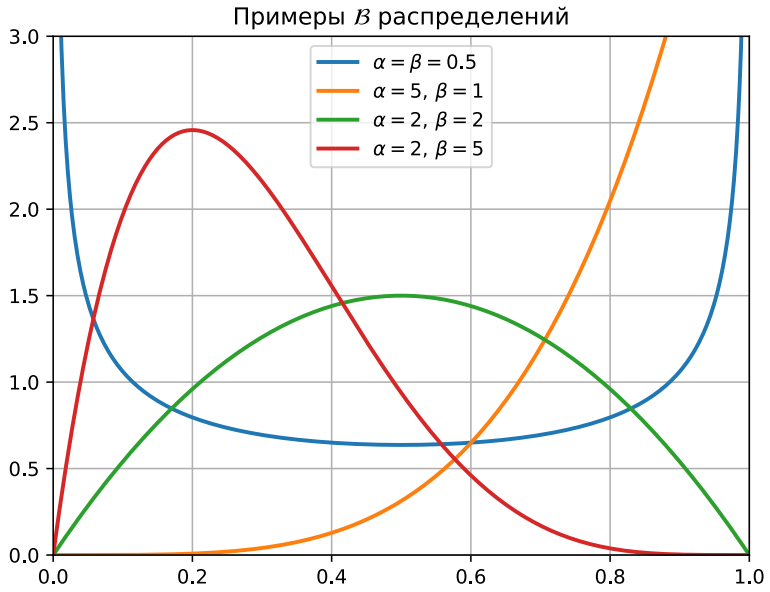

Рис. 6: Примеры $\mathcal{B}$ распределений

7.2.2 Калибровка Isotonic regression

Строится монотонно неубывающая кусочно-постоянная функция деформации оценок алгоритма (рис. 7). Требуется восстановить функцию $g(q(x))$, параметрами которой является измельчение сегмента $[0, 1]$, на каждом сегменте измельчения $g(t) = \texttt{const}$, $g(t_1) \leq g(t_2)$, $t_1 \leq t_2$, что в общем случае решается специальными солверами (например pool adjacent violators algorithm за $O(n)$).

$$\sum_{i=1}^{N} \left(k_i - g(q(x_i))\right)^2 \longrightarrow \min_g$$

7.2.3 Гистограммная калибровка

Является частным случаем Изотонической при фиксированных равных интервалах $\mathbb{B}_i$. Такой метод предсказывает лишь конечное число вероятностей, являющееся гиперпараметром алгоритма.

$$\sum_{i=1}^{\ell} \left| \frac{1}{\mathbb{B}_i} \sum_{j=1}^{N} \mathbf{1}_{\{q(x_i) \in \mathbb{B}_i\}} k_j - \theta_i \right| \longrightarrow \min_{\vec{\theta}}$$

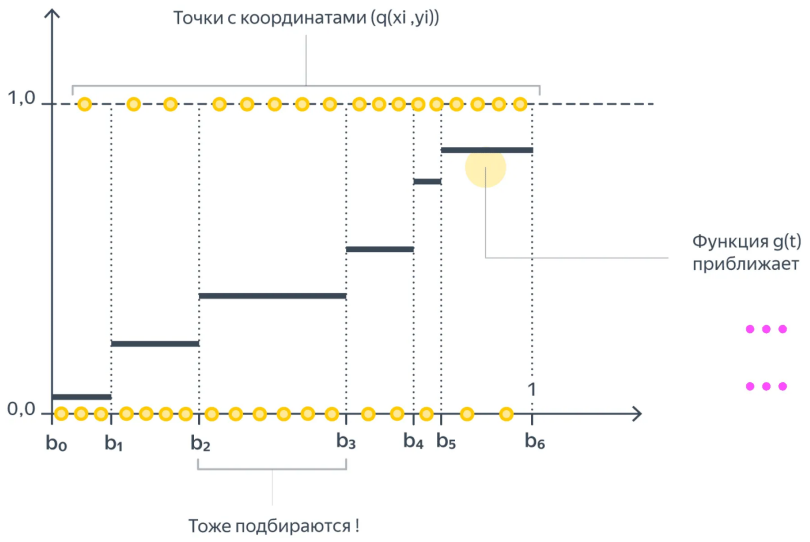

Рис. 7: Определение Изотонической калибровки

8 GLM

Теперь тонкий лед: вы будете проходить всю дальнейшую теорию на 4 курсе на Байесовских методах машинного обучения, поэтому все дальнейшие рассуждения будут носить лишь ознакомительный характер.

Модель GLM уже встречалась вам на лекциях Воронцова, ее суть заключается в описании семейства вероятностных стратеги через использование распределений **экспоненциального семейства** (с распределениями этого семейства каждый из вас встречался на курсе Теории вероятности и Математической статистики). В сущности, исследователь исходя из своих первичных знаний о задаче строит гипотезу о том, как выглядит распределение $p(k \mid x)$ (в частности как была разобрана logreg, там было Бернуллевское распределение), после чего выводит зависимости параметров распределения от наблюдения.

Общий вид экспоненциального распределения: $p(x \mid \theta) = \frac{f(x)}{g(\theta)} \exp\{\theta^T \vec{u}(x)\}$. Далее окажется, что наши модели стремятся приблизить именно θ – вектор параметров распределения по наблюдению. Однако сейчас мы рассмотрим частное представление семейства, что должно быть вам знакомо:

$$p(k_i \mid \theta_i, \varphi_i) = \exp\left\{\frac{k_i\theta_i - c(\theta_i)}{\varphi_i} + h(\theta_i, \varphi_i)\right\}$$

Оказывается, что $\mu_i = \mathbb{E}k_i = c'(\theta_i) \Longrightarrow \theta_i = (c')^{-1}(\mathbb{E}k_i) = \mathrm{g}(\mu_i); \mathbb{D}k_i =$

$\varphi_i c''(\theta_i)$ данное выражение в полной мере вы пощупаете в следующем году, нужно лишь сказать, что данное выражение является частным от много более общей теории, с которой в данном курсе мы не встречаемся.

8.1 Бернуллевское распределение

$$p(k_i \,|\, \mu_i) = \mu^{k_i}(1-\mu_i)^{1-k_i} = \{\text{приведем к явному виду}\} = \exp\left\{ k_i \underbrace{\ln \frac{\mu_i}{1-\mu_i}}_{\theta_i = logit} + \ln(1-\mu_i) \right\}$$

Значит параметр θ_i есть совсем не сама вероятность μ_i, а некоторая функция от нее: $\theta_i = \mathrm{g}(\mu_i) = \ln \frac{\mu_i}{1-\mu_i}$. Следовательно $\mu_i = \mathrm{g}^{-1}(\theta_i) = \frac{1}{1+e^{\theta_i}} = \sigma(\theta_i)$! Таким образом логистическая регрессия восстанавливает θ по которой мы уже и находим параметр вероятности.

8.2 Гауссовское распределение

$$p(k_i \,|\, \mu_i, \sigma_i^2) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left\{ -\frac{(k_i - \mu_i)^2}{2\sigma_i^2} \right\} = \exp\left\{ \frac{k_i\mu_i + -0.5\mu_i^2}{\sigma_i^2} - \frac{k_i^2}{2\sigma_i^2} - 0.5\ln(2\pi\sigma_i^2) \right\}$$

$$\Longrightarrow \theta_i = \mu_i, \quad \varphi_i = \sigma_i^2$$

Получив формулы оценок на параметры строим модели, что по объекту предсказывают θ_i и φ_i и подставляем в функционал эмпирического риска для оптимизации.

1. https://github.com/mmp-mmro-team/mmp_mmro_fall_2024/blob/main/seminars/Seminar_5_probabalistic_ml.pdf
2. Десять лекций по статистическому и структурному распознаванию образов. Михаил Иванович Шлезингер, Вацлав Главач
3. https://education.yandex.ru/handbook/ml/article/kak-ocenivat-veroyatnosti
4. Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration