Урок 13

Теорема Байеса в машинном обучении

13.1. Теорема Байеса

13.1.1. Ключевые понятия из теории вероятности

Пусть x и y — взаимозависимые случайные величины, тогда условная плотность на y при условии x по определению равна отношению совместной плотности распределения $p(x, y)$ к безусловной, или маргинальной, плотности $p(x)$:

$$p(y|x) = \frac{p(x, y)}{p(x)}.$$

Тогда:

$$p(y|x)p(x) = p(x, y) = p(x|y)p(y).$$

Отсюда следует формула обращения условной плотности:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}.$$

Если взять интеграл от обоих частей этого выражения:

$$1 = \int p(x|y) = \int \frac{p(y|x)p(x)}{p(y)} dx = \frac{1}{p(y)} \int p(y|x)p(x)dx.$$

Отсюда следует, что безусловная (маргинальная) плотность распределения на y:

$$p(y) = \int p(y|x)p(x)dx$$

Это свойство часто называют правилом суммирования вероятностей.

Из правила суммирования вероятностей и правила произведения вероятностей следует знаменитая теорема Байеса:

$$p(y|x) = \frac{p(x|y)p(y)}{\int p(x|y)p(y)dy}.$$

Теорема Баейса позволяет переходить от априорных распределений на неизвестную величину (в данном случае на y) к апостериорным распределениями на y при условии x:

$$Posterior = \frac{Likehood \times Prior}{Evidence}$$

13.1.2. Геометрический смысл условного и безусловного распределений

Рассматривается двухмерная случайная величина (ее плотность распределения $p(x, y)$ изображена на графике), компоненты которой взаимонезависимы.

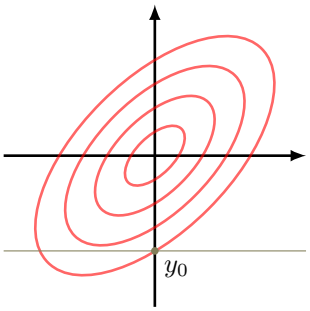

Рис. 13.1: Линии плотности случайной величины

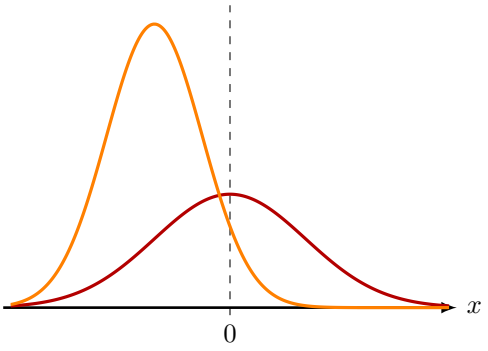

Рис. 13.2: Графики функций $p(x)$ (красный) и $p(x|y_0)$ (оранжевый).

Безусловное распределение $p(x)$ выражается как

$$p(x) = \int p(x, y)dy = \int p(x|y)p(y)dy$$

и является проекцией двухмерного графика на ось x. Очевидно, математическое ожидание x будет совпадать с матожиданием x в совместном вероятностном распределении. Если теперь стало известным значение $y = y_0$, то можно посчитать условное распределение на x при условии y_0:

$$p(x|y_0) = \frac{p(x, y_0)}{p(y_0)},$$

этому распределению будут отвечать сечения двухмерного графика гиперплоскостью $y = y_0$. Следует обратить внимание, что у распределения $p(x|y_0)$, по сравнению с безусловным $p(x)$, изменилось математическое ожидание и уменьшилась дисперсия.

13.1.3. Метод максимального правдоподобия

Ну и, наконец, давайте вспомним ключевой метод статистического оценивания из классической статистики, известный как метод максимального правдоподобия.

Стандартная задача матстатистики, как известно, — оценить значение неизвестных параметров распределения по заданной выборке из этого распределения.

Пусть $X = (x_1, ..., x_n)$ — выборка из независимых одинаково распределенных случайных величин:

$$x_i \sim p(x|\theta).$$

Плотность распределения $p(x|\theta)$ известна с точностью до параметров θ, которые необходимо оценить по известной выборке.

Метод максимального правдоподобия заключается в следующем. Сначала составляется функция правдоподобия выборки:

$$p(X, \theta) = \prod_{i=1}^{n} p(x_i|\theta).$$

Значения x_i являются известными, а значения θ выбираются таковыми, которые добавляют максимум функции правдоподобия:

$$\theta_{ML} = \operatorname{argmax}_\theta p(X, \theta) = \operatorname{argmax}_\theta \prod_{i=1}^{n} p(x_i|\theta) = \operatorname{argmax}_\theta \sum_{i=1}^{n} \ln p(x_i|\theta).$$

На практике обычно максимизируют не саму функцию правдоподобия, а ее логарифм, так как в этом случае произведение переходит в сумму логарифмов.

Для метода правдоподобия получен ряд интересных теоретических результатов:

- Оценка максимального правдоподобия является асимптотически несмещенной:

$$\mathbb{E}\theta_{ML} = \theta_* \qquad \text{при} \qquad n \gg 1.$$

- Оценка максимального правдоподобия является состоятельной:

$$\theta_{ML} \to \theta_* \qquad \text{при} \qquad n \to +\infty.$$

- Оценка максимального правдоподобия среди прочих несмещенных оценок обладает наименьшей возможной дисперсией, то есть обладает свйоством эффективности.

- Оценки максимального правдоподобия асимптотически нормальны при больших n, то есть имеют нормальное распределение с математическим ожиданием, равным истинному значению, и ковариационной матрицей, связанной с матрицей информации Фишера.

Тем не менее большинство теоретических результатов, которые гарантируют корректность и оптимальность оценивания по методу максимального правдоподобия, получены при условии $\frac{n}{d} \gg 1$, где n — количество объектов в выборке, d — размерность θ. Если это условие не выполняется, многие результаты метода максимального правдоподобия перестают быть корректными.

13.2. Байесовский подход к теории вероятностей

13.2.1. Отличие байесовского подхода от классического

На самом деле ключевое различие между частотным подходом, который изучается в вузах, и байесовским подходом заключается в том, как трактовать случайность.

	Частотный	**Байесовский**
Интерпретация случайности	Объективная неопределенность	Субъективное незнание
Метод вывода	Метод максимального правдоподобия	Теорема Байеса
Оценки	θ_{ML}	$p(\theta\|X)$
Применимость	$\frac{n}{d} \gg 1$, где d — размерность θ.	Любые n

Таблица: Сравнение байесовского и классического подходов.

С точки зрения классического подхода, случайная величина — это величина, значение которой мы принципиально не можем предсказать, то есть некоторая объективная неопределенность.

В то же время с точки зрения байесовского подхода случайная величина на самом деле является детерминированным процессом, просто часть факторов, которые определяют исход этого процесса, для нас неизвестны. Именно поэтому мы и не можем предсказать конкретный исход данного испытания с данной случайной величиной.

Из этого сразу вытекают некоторые следствия. Например, с точки зрения байесовского подхода любую неизвестную величину можно интерпретировать как случайную и использовать аппарат теории вероятности, в частности, вводить на нее плотность распределения. При этом, коль скоро случайные величины кодируют субъективное незнание, у разных людей неопределенность на одну и ту же случайную величину может быть разная. Именно поэтому и плотности распределения на эту случайную величину будут отличаться для разных людей, обладающих разной информацией о факторах, влияющих на эту случайную величину.

С точки зрения классического подхода величины четко делятся на случайную и детерминированную и бесмысленно применять аппарат теории вероятности к детерминированным случайным величинам или параметрам. С точки зрения байесовского подхода все величины, значения которых неизвестны, можно интерпретировать как случайные и, соответственно, можно вводить плотность распределения и выполнять байесовский вывод.

Основным методом оценивания в классическом подходе является метод максимального правдоподобия. При байесовском подходе к статистике основным выводом является теорема Байеса. Соответственно, результатом оценивания в классическом подходе обычно являются точечные оценки, как правило, это оценки

максимального правдоподобия, либо реже — доверительные интервалы. При байесовском же подходе результатом вывода является апостериорное распределение на оцениваемые параметры. Метод максимального правдоподобия является оптимальным при $n \to \infty$, соответственно, большинство теорем в теории вероятности, которые обосновывают корректность применения этого метода, доказывают предположение, что объем выборки, по которой мы оцениваем неизвестный параметр, много больше 1.

В то же время байесовский подход можно использовать при любом объеме выборки, даже если объем выборки равен 0. В этом случае результатом байесовского вывода и апостериорного распределения просто будет являться априорное распределение. В то же время, если объем выборки, а именно отношение n к d, где n — это количество объектов, а d — это размерность оцениваемых параметров, много больше 1, результат байесовского вывода начинает стремиться к результату, оцениваемому с помощью метода максимального правдоподобия. Тем самым все теоретические гарантии, которые известны для метода максимального правдоподобия, применимы и к результату байесовского вывода.

13.2.2. Иллюстративный пример

Одним из преимуществ байесовского подхода является возможность объединения разных вероятностных моделей, которые отражают те или иные косвенные характеристики оцениваемой неизвестной величины.

Пусть $y_1, ..., y_m$ — несколько косвенных проявлений неизвестной величины x, для каждого из которых существует вероятностная модель $p_j(y_j|x)$, определяющая вероятность наблюдения того или иного значения y_j. Необходимо оценить x путем объединения информации из $y_1, ..., y_m$.

Пусть $p(x)$ выражает исходные представления о возможных значениях x. Тогда после применения формулы Байеса можно получить апостериорное распределение на x при условии $y = y_1$:

$$p(x|y_1) = \frac{p_1(y_1|x)p(x)}{\int p_1(y_1|x)p(x)dx}.$$

При анализе результата второго измерения, которое может быть никак не связано с первым измерением и получено из совершенно другой вероятностной модели, нужно снова применить байесовский вывод, только теперь в качестве априорного распределения на x мы положим апостериорное распределение, полученное после измерения y_1:

$$p(x|y_1, y_2) = \frac{p_2(y_2|x)p(x|y_1)}{\int p_2(y_2|x)p(x|y_1)dx}.$$

Действуя так m раз, можно получить апостериорное распределение на x при условии $y_1, ..., y_m$, которое отражает максимум информации, которую можно извлечь в данном случае:

$$p(x|y_1, ..., y_m) = \frac{p_m(y_m|x)p(x|y_1, ..., y_{m-1})}{\int p_m(y_m|x)p(x|y_1, ..., y_{m-1})dx}.$$

Если бы вместо апостериорных распределений использовались точечные оценки, это было бы похоже на положении слепых мудрецов из известной притчи, которые пытались изучать слона на основе различных тактильных ощущений. Как известно, в притче мудрецы не смогли прийти к единому мнению, в то же время, если бы они оперировали байесовским аппаратом и получали бы апостериорное распределение, скорее всего, они смогли бы прийти к мнению относительно того, что же они изучают.

13.3. Байесовские модели в задачах машинного обучения

Задачу машинного обучения можно интерпретировать как задачу восстановления зависимости между наблюдаемыми и скрытыми компонентами. При это зависимость восстанавливается по обучающей выборке, в которой предполагается, что мы знаем и наблюдаемые, и скрытые компоненты.

Если оперировать в рамках вероятностного подхода или, более конкретно, в рамках байесовского подхода, то в качестве модели зависимости между наблюдаемыми и скрытыми компонентами используются совместные вероятностные распределения над наблюдаемыми, скрытыми компонентами и, если речь идет про байесовский подход, то еще и над параметрами, которые настраиваются в ходе процедуры обучения.

13.3.1. Задача линейной регрессии

В задаче линейной регрессии есть три группы переменных:

- $x \in \mathbb{R}^d$ — признаки объекта (наблюдаемые параметры),
- $t \in \mathbb{R}$ — целевая переменная (скрытые параметры),
- $w \in \mathbb{R}^d$ — веса линейной регрессии (эти параметры настраиваются в ходе процедуры обучения).

Тогда на байесовском языке такую модель линейной регрессии можно сформулировать в виде совместного вероятностного распределения $p(t, w|x)$ на t и w при условии x. Такие модели называются дискриминативными моделями. В противовес им рассматривают так называемые генеративные модели, в которых совместные распределения вводятся как на скрытые величины t и настраиваемые параметры w, так и на наблюдаемые параметры x. В рамках генеративных моделей помимо прочего возможно решать задачу по генерации новых объектов. В рамках же дискриминативных моделей класс задач, которые можно решать, ограничивается задачами прогноза скрытой компоненты t при условии наблюдаемой компоненты x.

В следующей дискриминативной модели:

$$p(t, w|x) = p(t, x|w)p(w) = \mathcal{N}(t|w^T x, \sigma^2)\mathcal{N}(w|0, I).$$

совместное распределение $p(t, w|x)$ было разложено по правилам произведения на функции правдоподобия $p(t, x|w)$ и априорное распределение $p(w)$ на w. В качестве функции правдоподобия и априорного распределения используются нормальные распределения с соответствующими параметрами.

Пусть задана обучающая выборка $(X, T) = (x_i, t_i)_{i=1}^n$ и выполняется поиск максимума апостериорного распределения:

$$\begin{aligned} w_{MP} &= \operatorname{argmax}_w p(w|X, T) = \operatorname{argmax}_w p(T|X, w)p(w) = \operatorname{argmax}_w \prod_{i=1}^n p(t_i|x_i, w)p(w) = \\ &= \operatorname{argmax}_w \left[\sum_{i=1}^n \ln p(t_i|x_i, w)p(w) + \ln p(w) \right] = \operatorname{argmin}_w \frac{1}{2\sigma^2} \sum_{i=1}^n (t_i - x_i^T w)^2 + \frac{1}{2}\|w\|^2 = \\ &= \operatorname{argmin}_w \left[\sum_{i=1}^n (t_i - x_i^T w)^2 + \sigma^2 \|w\|^2 \right]. \end{aligned}$$

Фактически был получен метод наименьших квадратов с L_2-регуляризатором. Таким образом, известный метод наименьших квадратов с L_2-регуляризатором может быть переформулирован на языке байесовских моделей и соответствует достаточно простой вероятностной модели, в которую было введено гауссовское априорное распределение с нулевым матожиданием на веса линейной регрессии.

13.3.2. Преимущества байесовских моделей в машинном обучении

Первое преимущество состоит в возможности построения сложных вероятностных моделей из более простых. Это становится возможным благодаря тому, что результат байесовского вывода в одной модели (то есть апостериорное распределение) можно использовать в качестве априорного распределения в следующей вероятностной модели. Тем самым происходит зацепление разных вероятностных моделей.

Еще одним преимуществом байесовских методов является возможность обработки массива данных, которые поступают последовательно. В самом деле, используя апостериорное распределение в качестве априорного при поступлении новой порции данных, можно легко произвести обновление апостериорного распределения без необходимости повторного обучения модели с нуля. При использовании точечных оценок нужно было бы заново обучать модель.

Еще одним преимуществом байесовских методов является возможность использования априорного распределения, которое предотвращает излишнюю настройку неизвестных параметров под обучающую выборку. Это в свою очередь позволяет избежать эффекта переобучения, который часто свойственен даже задачам, в которых присутствует гигантский объем обучающих выборок, но в ситуации, когда и количество настраиваемых параметров тоже достаточно велико. Благодаря использованию априорных распределений можно регуляризовывать модель машинного обучения и предотвращать эффект переобучения.

Наконец, одним из ключевых достоинств байесовских методов является возможность работы с не полностью размеченными, частично размеченными, а то и вовсе не размеченными обучающими выборками. То есть в ситуациях, когда в обучающих выборках известна наблюдаемая компонента, а скрытая компонента известна не для всех объектов, либо для многих объектов известно не точное значение скрытой компоненты, а лишь некоторое допустимое подмножество скрытых компонент. Такие выборки называются частично размеченными. Оказывается, что байесовский формализм, байесовское моделирование позволяет абсолютно корректно работать с такими моделями и извлекать из них максимум имеющейся информации о неизвестных значениях параметров.