Студенческая научная группа "Факторный анализ и прогнозирование"

Байесовские методы. ЕМ-алгоритм. Общие сведения

Зеленин Герман

Московский государственный университет, механико-математический факультет

12 ноября 2022

Содержание

Основные понятия

Definition 1

Пусть x и y - две случайные величины. Тогда **условным распределением** $p(x|y)$ x относительно y называется отношение *совместного распределения* $p(x, y)$ и *маргинального распределения* $p(x)$:

$$p(x|y) = \frac{p(x, y)}{p(y)} \tag{1}$$

Основные понятия

Theorem 2

Пусть $x_1, x_2, ..., x_n$ - случайные величины. Тогда их совместное распределение можно представить в виде произведения n одномерных условных распределений с постоянно уменьшающейся посылкой:

$$p(x_1, ..., x_n) = p(x_n|x_1, ..., x_{n-1})...p(x_2|x_1)p(x_1) \tag{2}$$

Основные понятия

Theorem 3

Пусть $x_1, x_2, ..., x_n$ *- случайные величины. Если известно их совместное распределение* $p(x_1, ..., x_n)$, *то совместное распределение подмножества случайных величин* $x_1, ..., x_k$ *будет равно:*

$$p(x_1, ..., x_k) = \int p(x_1, ..., x_n) dx_{k+1} ... dx_n \tag{3}$$

Основные понятия

Definition 4

Будем говорить, что распределение $p(x|\theta)$ лежит в **экспоненциальном классе**, если оно может быть представлено в следующем виде:

$$p(x|\theta) = \frac{f(x)}{g(\theta)} exp(\theta^T u(x)) \tag{4}$$

Содержание

Частотный подход. Метод максимального правдоподобия

Рассмотрим некоторую выборку $X = (X_1, X_2, ..., X_n)$ из некоторого параметрического распределения $p_\theta(x)$

Задача: хотим оценить параметр θ так, чтобы вероятность пронаблюдать то, что мы пронаблюдали, была максимальна.

$$\theta_{ML} = \arg\max_\theta p(X|\theta) = \arg\max_\theta \prod_{i=1}^{n} p(x_i|\theta) = \arg\max_\theta \sum_{i=1}^{n} \log p(x_i|\theta) \tag{5}$$

Частотный подход. Метод максимального правдоподобия

Во многих частных случаях сумма логарифмов правдоподобий будет выпуклой вверх функцией, то есть у неё один максимум, который достаточно легко найти даже в пространствах высокой размерности.

Заметим, что θ_{ML} - случайная величина, так как является функцией от выборки.

Свойства оценки максимума правдоподобия

1. Состоятельность: ОМП сходится к истинному значению параметров по вероятности при $n \to +\infty$,
2. Асимптотически несмещенная: $\theta_{ML} = E[\theta]$ при $n \to +\infty$,
3. Асимптотически нормальная: θ_{ML} распределена нормально при $n \to +\infty$
4. Асимптотическая эффективность: ОМП обладает наименьшей дисперсией среди всех состоятельных асимптотически нормальных оценок.

Чем частотный подход плох в машинном обучении?

Как уже было сказано, в методе максимального правдоподобия мы выбираем параметр распределения так, чтобы вероятность пронаблюдать то, что мы пронаблюдали была максимальной. Говоря на языке машинного обучения, мы подгоняем параметры распределения под полученные данные, а это чревато переобучением.

Байесовский подход. Теорема Байеса

Theorem 5

Пусть x и y - случайные величины. Тогда

$$p(x|y) = \frac{p(x|y)p(y)}{\int p(x|y)p(y)dy} \tag{6}$$

где $p(y|x)$ - апостериорное распределение, $p(x|y)$ - правдоподобие, $p(y)$ - априорное распределение.

Байесовский подход

Пусть у нас есть априорное распределение $p(\theta)$, которое отражает некую внешнюю информацию о возможных значениях параметров (если такой информации нет, мы всегда можем ввести неинформативное распределение). Тогда результатом применения теоремы Байеса будет апостериорное распределение на параметры:

$$p(\theta|X) = \frac{\prod p(x_i|\theta)p(\theta)}{\int \prod p(x_i|\theta)p(\theta)d\theta} \tag{7}$$

Ответом является новое распределение на параметры модели, в отличии от метода максимального правдоподобия, где ответом являлось конкретное значение параметров.

Содержание

EM-алгоритм. Формулировка решаемой задачи

Задача 1: По выборке X восстановить параметры θ распределения методом максимального правдоподобия:

$$p(X|\theta) \to \max_{\theta} \tag{8}$$

Вопрос: В каких параметрических семействах эту задачу можно решать эффективно?

EM-алгоритм. Формулировка решаемой задачи

Ответ: Если плотность распределения $p(X|\theta)$ лежит в экспоненциальном классе, то мы можем эффективно найти оценку максимального правдоподобия для параметров θ. Иногда это возможно в явном виде (дифференцируем логарифм правдоподобия, приравниваем к нулю, и находим из полученной системы уравнений параметры θ), а в остальных случаях можно построить эффективную численную процедуру оценки (благодаря тому, что логарифм функции правдоподобия — вогнутая функция).

EM-алгоритм. Формулировка решаемой задачи

Проблема заключается в том, что экспоненциальный класс не такой широкий, как могло бы показаться. Зачастую на практике наблюдаемые данные имеют гораздо более сложное распределение, которое в экспоненциальный класс никак не вписывается.

Пример: смесь гауссиан

Наши данные пришли из сложного распределения. Но если бы мы знали что-нибудь еще, то наше распределение стало бы куда более простым.

EM-алгоритм. Формулировка решаемой задачи

Задача 2: Введем латентные переменные Z так, чтобы совместное распределение $p(X, Z|\theta)$ лежало в экспоненциальном классе. Вместо предыдущей задачи, будем решать следующую задачу:

$$p(X, Z|\theta) \to \max_{\theta} \tag{9}$$

Замечание: Помимо нахождения решения исходной задачи, мы найдем так же информацию по латентным переменным.

Вывод ЕМ-алгоритма

Записываем цепочку преобразований:

$$\begin{aligned}\log p(X|\theta) = \int q(Z) \log p(X|\theta) dZ = \\ = \int \log \frac{p(X,Z|\theta)}{p(Z|X,\theta)} dZ = \int \log \frac{p(X,Z|\theta) q(Z)}{p(Z|X,\theta) q(Z)} dZ = \\ = \int q(Z) \log \frac{p(X,Z|\theta)}{q(Z)} dZ + \int q(Z) \log \frac{q(Z)}{p(Z|X,\theta)} dZ \end{aligned} \tag{10}$$

где $p(q)$ - произвольное распределение в пространстве латентных переменных.

Дивергенция Кульбака - Лейбера

Definition 6

Дивергенция Кульбака-Лейбера между двумя распределениями p и q определяется следующим образом:

$$KL(p(x)||q(x)) = \int p(x) \log \frac{p(x)}{q(x)} dx \tag{11}$$

Theorem 7

$KL(p||q) \geqslant 0$, причем $KL(p||q) = 0$ если и только если эти распределения почти всюду (везде кроме множества меры ноль) совпадают.

Вывод EM-алгоритма

Заметим, что в выражении (10) второе выражение является KL-дивергенцией распределений $q(Z)$ и $p(Z|X,\theta)$. Так как она неотрицательна, то можем записать следующее неравенство:

$$\log p(X|\theta) \geqslant \int q(Z) \log \frac{p(X,Z|\theta)}{q(Z)} dZ \tag{12}$$

Идея EM-алгоритма состоит в том, чтобы вместо оптимизации логарифма неполного правдоподобия оптимизировать полученную нижнюю оценку, но теперь уже как по θ так и по распределению q

Вывод EM-алгоритма

Definition 8

Правая часть выражения 12 называется *нижней границей на обоснованность (ELBO)* и обозначается как $\mathcal{L}(q, \theta)$

Нижняя оценка на обоснованность является вариационной нижней оценкой, то есть удовлетворяет тому, что всегда не превосходит выражения, которое оценивает (что как раз говорит выражение 12), а так же для любого аргумента исходной функции (θ) найдутся такие значения вариационных (q), для которых неравенство превращается в равенство.

Вывод ЕМ-алгоритма

Воспользуемся этими свойствами и перейдем от оптимизации неполного правдоподобия к оптимизации нижней оценки на обоснованность. Будем решать данную задачу итерационно:

1) Оптимизировать по q при фиксированном θ (Е-шаг):

$$\mathcal{L}(q, \theta_0) \longrightarrow \max_q \Rightarrow q(Z) = p(Z|X, \theta) \tag{13}$$

2) Оптимизировать по θ при фиксированном q (М-шаг):

$$\mathcal{L}(q_0, \theta) \longrightarrow \max_\theta \Leftrightarrow \int q(Z) \log p(X, Z|\theta) dZ \longrightarrow \max_\theta \tag{14}$$

Вывод EM-алгоритма

На E-шаге задача функциональной оптимизации. Сумма в (10) не зависит от q, а потому максимизация по q первого слагаемого эквивалентна минимизации по q второго, а второе слагаемое - KL-дивиргенция. Мы знаем, что она достигает минимума, следовательно, приравниваем $q(Z) = p(Z|X, \theta)$. Если можем найти апостериорное распределение $p(Z|X, \theta)$, то E-шаг проделывается в явном виде.

Содержание

Вывод ЕМ-алгоритма

1) Разделение смесей распределения (латентные переменные - номера распределений, из которых пришли данные).
2) Метод главных компонент - метод уменьшения размерности данных, потеряв наименьшую информацию.
3) Задачи классификации.