Теория вероятности

Дима Трушин

Семинар 6

Задача математической статистики

В теории вероятностей, изучая какой-нибудь черный ящик, всегда предполагается, что за ним стоит некоторое вероятностное пространство (Ω, P). Кроме того, обычно к черному ящику цепляют какой-нибудь измерительный прибор, то есть рассматривают случайную величину $\xi\colon \Omega \to \mathbb{R}$. В этом случае главный интерес для изучения – это мера P_ξ, которая полностью описывает все свойства прибора ξ на нашем черном ящике. Типичная задача теории вероятности звучит так: мы знаем меру P_ξ, предсказать, как будет отвечать прибор на наши запросы. В математической статистике по сути решается обратная задача. Мы не знаем как устроено (Ω, P), мы понятия не имеем, как выглядит мера P_ξ, но мы знаем, что прибор уже успел дать несколько ответов $x_1, \ldots, x_n \in \mathbb{R}$. Задача – вытащить как можно больше информации про неизвестные компоненты черного ящика. Как не трудно догадаться, раз мы опрашиваем не сам черный ящик, а лишь висящий на нем прибор, то (Ω, P) находится вне нашей досягаемости. Самое лучше, что мы можем попытать узнать – это как устроена случайная величина $\xi\colon \Omega \to \mathbb{R}$. Но раз мы не знаем, что из себя представляет Ω, лучшее, что мы можем узнать – это как устроена мера P_ξ. С другой стороны, если мы знаем вероятностную меру P_ξ для случайной величины ξ, то мы знаем про нее все. Таким образом общая задача математической статистики выглядит так.

Задача Пусть на некотором вероятностном пространстве (Ω, P) задана случайная величина $\xi\colon \Omega \to \mathbb{R}$ и пусть заданы измерения случайной величины $x_1, \ldots, x_n \in \mathbb{R}$, мы хотим восстановить P_ξ или какие-то другие характеристики ξ.

Как надо думать про измерения? Когда мы замеряли, скажем, значение x_1, наш черный ящик находился в каком-то состоянии ω_1, потому $x_1 = \xi(\omega_1)$. Потом, когда мы измеряли x_2, наш черный ящик находился в состоянии ω_2, то есть $x_2 = \xi(\omega_2)$ и т.д. Таким образом, можно думать, что $x_1, \ldots, x_n$ – это последовательность $\xi(\omega_1), \ldots, \xi(\omega_n)$. Значит про эту задачу можно думать, как про задачу восстановления функции $\xi\colon \Omega \to \mathbb{R}$ по значению в конечном числе точек. Однако, такая модель не самая лучшая. Основная проблема в том, что мы не знаем, как именно прибор переключается между состояниями $\omega_1, \ldots, \omega_n$, где именно в пространстве Ω они сгруппированы. Кроме того, не ясно, влияет ли первое измерение на второе и если влияет, то как учитывать подобное в данной модели? По-хорошему эту модель надо дополнить некоторой информацией о переходе из одного состояния в другое при замерах. Но это автоматически означает, что нам надо знать какие есть состояния у черного ящика, чтобы описать подобный механизм. А такая информация может быть просто недоступна нам. Оказывается, можно закодировать эту информацию по-другому, так что необходимость знать Ω полностью пропадает.

Основная модель математической статистики

Как и выше, мы должны предположить, что у нас есть некоторый изучаемый черный ящик, за которым стоит некоторое неизвестное нам вероятностное пространство (Ω, P). Разница будет в том, как мы будем моделировать наш измеряющий прибор.

Предположим, что мы взяли некоторый прибор $\xi\colon \Omega \to \mathbb{R}$ и начинаем делать измерения. Сначала сделали первое измерение, получили x_1. Потом взяли тот же самый эксперимент повторили с нуля и получается независимо сделали второе измерение и получили x_2 и т.д. На этот процесс можно смотреть вот каким образом. Мы можем использовать ξ как эталонный прибор, а для n независимых измерений сделать n его независимых копий: $\xi_1, \ldots, \xi_n\colon \Omega \to \mathbb{R}$. Тогда вместо того, чтобы делать n измерений подряд с помощью одного прибора, мы просто разом сделаем n измерений с помощью n независимых приборов. То есть мы можем считать, что

$$x_1 = \xi_1(\omega), x_2 = \xi_2(\omega), \ldots, x_n = \xi_n(\omega)$$

В таком подходе мы думаем, что за каждое измерение отвечает своя случайная величина. Но все эти величины являются копиями одной величины ξ. Математически это означает, что $P_{\xi_i} = P_\xi$. Заметим, что мы при описании процесса предполагали, что будем делать измерения независимыми друг от друга. Это означает, что все случайные величины $\xi_1, \ldots, \xi_n$ должны быть в совокупности независимыми. Однако теперь, когда за каждое измерение отвечает своя случайная величина, мы можем (если захотим) отбросить условие независимости и считать, что они могут быть зависимыми. Кроме того за свойство зависимости или независимости будет отвечать совместное распределение $P_{(\xi_1,\ldots,\xi_n)}$ в $\mathbb{R}^n$. Таким образом мы избавились от необходимости знать Ω, чтобы понять, какая взаимосвязь между измерениями. Я бы еще добавил, что предположение о независимости измерений часто является очень естественным и потому популярно в приложениях. Один из самых популярных способов описания зависимостей являются Марковские цепи, которых мы касаться не будем, но слышать эти слова надо. Давайте теперь формально опишем задачу и основную модель математической статистики.

Задача Пусть на некотором вероятностном пространстве (Ω, P) заданы случайные величины $\xi_1, \ldots, \xi_n$ такие, что они одинаково распределены (то есть $P_{\xi_i} = P_{\xi_j}$) и, например, независимы. Пусть нам задан результат измерений

$$x_1 = \xi_1(\omega), \ldots, x_n = \xi_n(\omega) \in \mathbb{R}$$

мы хотим восстановить $P_\xi = P_{\xi_i}$ на прямой или какие-то численные характеристики этой вероятностной меры.

Замечания

1. Чтобы описать семейство одинаково распределенных случайных величин обычно делают так. Говорят, пусть ξ – некоторая случайная величина с интересующем нас распределением P_ξ. А после этого говорят, что пусть нам заданы случайные величины $\xi_1, \ldots, \xi_n$ распределенные так же как и ξ, то есть $P_{\xi_i} = P_\xi$. Обычно этот факт пишут так $\xi_i \sim \xi$, читается «ξ_i распределена так же как ξ». И тогда говорят уже про восстановление распределения для ξ или характеристик распределения для ξ.

2. Мы во всех задачах (но не во всех) в дальнейшем будем предполагать, что случайные величины $\xi_1, \ldots, \xi_n$ независимы в совокупности. Это одна из самых популярных гипотез и она очень сильно упрощает рассуждения, мы же не хотим себя мучить.

3. В подобных задачах, часто предполагается из каких-нибудь теоретических соображений, что ξ вообще говоря распределена не как попало, а подчиняется какому-нибудь закону, у которого нам надо восстановить параметры. Например, может быть известно, что ξ – дискретная случайная величина, принимающая 0 и 1, но мы не знаем с какими вероятностями и их надо восстановить. Или может быть известно, что ξ – непрерывная случайная величина и ее плотность имеет вид $p(x, \theta_1, \ldots, \theta_n)$, где θ_i – какие-то настраиваемые параметры, и нам надо установить какие эти параметры были на самом деле. Скажем, может оказаться, что ξ является нормально распределенной случайной величиной с неизвестными параметрами матожидания и дисперсии и нам надо их определить.

Оценки

Пусть ξ – некоторая случайная величина и $x_1, \ldots, x_n \in \mathbb{R}$ – ее независимые измерения. Мы хотим восстановить математическое ожидание и дисперсию случайной величины ξ. Для того, чтобы решить эту задачу, нам надо написать какую-нибудь формулу $\mathbb{E}\xi = \phi(x_1, \ldots, x_n)$. Вообще говоря, это может быть абсолютно любое выражение от переменных $x_1, \ldots, x_n$. Подобная формула $\phi(x_1, \ldots, x_n)$ называется оценкой (в данном случае оценкой математического ожидания). Однако, если мы напишем совсем дурацкую формулу, например, $\phi(x_1, \ldots, x_n) = 0$, то получим дурацкий ответ, ничего не говорящий о нашей случайной величине, а хотелось бы что-то про нее да узнать. Для этого есть несколько характеристик, которые можно проверить у оценки. Давайте разберем все эти премудрости на примере математического ожидания и дисперсии.

Выборочное математическое ожидание Самая популярная формула для оценки математического ожидания – среднеарифметическое из измеренных значений. А именно

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n}$$

Чтобы понять на сколько это хорошая формула, давайте вспомним, что в рамках нашей модели, мы считаем, что

$$x_1 = \xi_1(\omega), \ldots, x_n = \xi_n(\omega)$$

Тогда наша оценка считает

$$\eta(\omega) = \frac{\xi_1(\omega) + \ldots + \xi_n(\omega)}{n}$$

Но мы не знаем, в какой именно точке ω мы были. Потому вообще говоря, то что мы получили будет случайная величина

$$\eta = \frac{\xi_1 + \ldots + \xi_n}{n}$$

Несмещенность Первый важный вопрос такой: а попадаем ли мы с помощью этой оценки в математическое ожидание хотя бы в среднем? То есть верно ли, что $\mathbb{E}\eta = \mathbb{E}\xi$? Если посчитать

$$\mathbb{E}\eta = \mathbb{E}\left(\frac{\xi_1 + \ldots + \xi_n}{n}\right) = \frac{\mathbb{E}\xi_1 + \ldots + \mathbb{E}\xi_n}{n} = \frac{\mathbb{E}\xi + \ldots + \mathbb{E}\xi}{n} = \mathbb{E}\xi$$

Подобная характеристика называется несмещенностью.

Виды сходимости Когда мы изучаем оценки вида $\phi(x_1, \ldots, x_n)$ то обычно у нас есть оценка для любого натурального n (или как минимум сколь угодно большого). Тогда интересным является такой вопрос: а к чему стремится выражение $\phi(\xi_1, \ldots, \xi_n)$ при $n \to \infty$? Чтобы на него ответить, для начала надо понять, а в каком смысле вообще случайные величины могут стремиться друг к другу. Оказывается, что существует несколько разных видов сходимости на случайных величинах, о них и пойдет сейчас речь.

Случайные величины – это функции на пространстве Ω. Потому если мы говорим о сходимости случайных величин, мы просто говорим о сходимости функций. Давайте перечислим самые популярные виды сходимости на функциях. Пусть $\xi, \xi_n \colon \Omega \to \mathbb{R}$ – случайные величины.

1. **Равномерная сходимость.** Последовательность ξ_n равномерно сходится к ξ, пишут $\xi_n \rightrightarrows \xi$, при $n \to \infty$, если $\sup_{\omega\in\Omega} |\xi_n(\omega) - \xi(\omega)| \to 0$.

   Это значит, что во всех элементарных исходах ξ_n стремится к ξ с некой гарантированной общей скоростью. Это самый сильный вид сходимости и одновременно самый бессмысленный для случайных величин, ибо вероятностное пространство Ω вообще говоря определено с точностью до события вероятности 0. Подробно об этом ниже.

2. **Поточечная сходимость (сходимость всюду).** Последовательность ξ_n сходится к ξ поточечно (всюду), пишут $\xi_n \to \xi$, если для любого $\omega \in \Omega$ последовательность $\xi_n(\omega)$ сходится к $\xi(\omega)$.

   Это самый простой вид сходимости и вряд ли нуждается в дополнительном пояснении. Мы просто хотим на любом исходе сходиться.

3. **Сходимость почти всюду (почти наверное).** Последовательность ξ_n сходится к ξ почти всюду (почти наверно), пишут $\xi_n \xrightarrow{\text{a.s.}} \xi$, если существует подмножество $\Omega_0 \subseteq \Omega$ такое, что $P(\Omega_0) = 1$ и для любого $\omega \in \Omega_0$ последовательность $\xi_n(\omega)$ сходится к $\xi(\omega)$.

   Это самый популярный вид сходимости. Дело в том, что в теории вероятности все определено с точностью до события вероятности ноль. Мы можем изменить вероятностное пространство на множество вероятности ноль и ничего не изменится. Потому формально мы не можем гарантировать ничего в отдельно взятом исходе. Потому поточечная сходимость (как и равномерная) не имеют особенного смысла для случайных величин. Чтобы исправить ситуацию, пользуются сходимостью почти всюду (или как принято в теории вероятностей почти наверно). Эта та же самая поточечная сходимость, но только на множестве вероятности 1. То есть последовательность ξ_n может не сходиться к ξ но только вероятность такого события равна нулю, формально $P(\{\omega \mid \xi_n(\omega) \not\to \xi(\omega)\}) = 0$.

4. **Сходимость в среднем.** Фиксируем число $1 < p < \infty$. Тогда ξ_n сходится в $L_p(\Omega)$ к ξ, пишут $\xi_n \xrightarrow{L_p} \xi$, если $\mathbb{E}\left((\xi_n - \xi)^p\right) = \int_\Omega (\xi_n - \xi)^p \, dP \to 0$. Самые популярные сходимости в среднем: сходимость в средне квадратичном, то есть в $L_2(\Omega)$ и сходимость в $L_1(\Omega)$.

   Про сходимость в среднем можно думать так: среднее от модуля разности (может быть в какой-то степени p) от случайных величин идет к нулю. То есть площадь между графиками функций ξ_n и ξ идет к нулю. То есть мы может быть и не сходимся поточечно, но зазор между графиками уменьшается.

5. **Сходимость по мере (по вероятности).** Последовательность ξ_n сходится к ξ по мере (по вероятности), пишут $\xi_n \xrightarrow{P} \xi$, если для любого $\varepsilon > 0$ последовательность $P(\{\omega \mid |\xi_n(\omega) - \xi(\omega)| > \varepsilon\})$ сходится к нулю.

   Неформально это означает, что мера тех точек, в которых ξ_n дальше от ξ на ε стремится к нулю. То есть как только n все больше и больше, у нас все больше и больше точек, в которых мы все ближе и ближе к пределу.

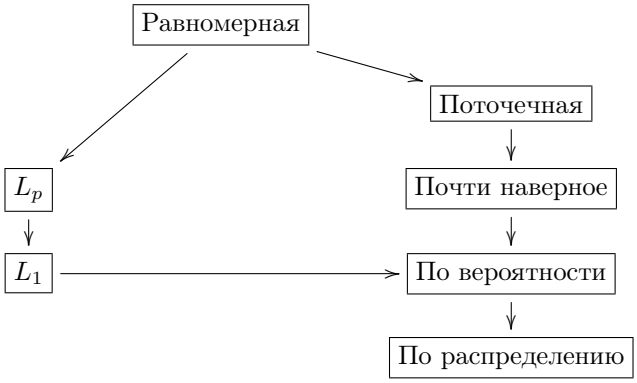
Сходимость по распределению. Кроме того, каждая случайная величина $\xi\colon \Omega \to \mathbb{R}$ дает меру P_ξ на прямой $\mathbb{R}$. Потому можно говорить о сходимости мер. Есть такая замечательная наука – функциональный анализ. Она придумала как можно проверять сходимость мер друг к другу. Этих видов сходимостей тоже очень много, но мы остановимся только на одной из них: $*$-слабая сходимость (сходимость по распределению).

Пусть ξ_n, ξ – случайные величины. Следующие условия эквивалентны:

1. Для любой непрерывной ограниченной функции $f\colon \mathbb{R} \to \mathbb{R}$ верно, что $\mathbb{E}(f(\xi_n)) \to \mathbb{E}(f(\xi))$.
2. Для любой точки $x \in \mathbb{R}$, в которой функция распределения $F_\xi(x)$ непрерывна, верно, что $F_{\xi_n}(x)$ сходится к $F_\xi(x)$.
3. Для любого отрезка $[a, b]$ такого, что $P(\xi = a) = P(\xi = b) = 0$, верно $P(\xi_n \in [a, b]) \to P(\xi \in [a, b])$.

В этом случае будем говорить, что последовательность ξ_n $*$-слабо сходится к ξ или что то же самое сходится по распределению и будем писать $\xi_n \xrightarrow{\mathcal{D}} \xi$. Важно понимать, что предел по распределению не единственный.

На диаграмме ниже показано, какие виды сходимостей какие влекут:



Сходимость и качество оценок Давайте еще раз посмотрим на нашу оценку

$$\bar{x} = \frac{\xi_1(\omega) + \ldots + \xi_n(\omega)}{n}$$

Теперь мы знаем, что в среднем мы попадем в математическое ожидание. Но вдруг для одних точек ω подобное среднее будет далеко от матожидания, а для других близко? Тогда мы вообще говоря не можем ничего гарантировать. Однако, тут на помощь приходят различные ассимптотические законы. Вот пример закона больших чисел:

Утверждение (Закон больших чисел)**.** *Пусть $\xi_1, \ldots, \xi_n, \ldots \sim \xi$ – последовательность одинаково распределенных независимых случайных величин. Тогда последовательность*

$$\frac{\xi_1 + \ldots + \xi_n}{n}$$

сходится к $\mathbb{E}\xi$ почти наверное.

Давайте расшифруем, что здесь сказано. Здесь говорится, что если я возьму почти любое $\omega \in \Omega$, то имеется сходимость

$$\frac{\xi_1(\omega) + \ldots + \xi_n(\omega)}{n} \to \mathbb{E}\xi, \quad n \to \infty$$

Фразу «почти любое» надо понимать так: пусть $X \subseteq \Omega$ – множество тех ω для которых есть такая сходимость, тогда $P(X) = 1$. Это значит, что если вероятность попасть в ω для которого такой сходимости нет равна нулю. На практике это означает, что мы всегда попадаем в ω, для которого выполнена подобная сходимость.

Таким образом, этот ассимптотический закон говорит о том, что не только в среднем наша формула дает математическое ожидание, но вообще всегда (формально почти всегда) результат по этой формуле будет сколь угодно близко стремиться к настоящему математическому ожиданию.

Надо сказать, что у этого закона есть куча разных вариаций. Есть вариация, которая даже оценивает скорость сходимости. А именно, формулировка приблизительного такого вида

$$P\left(\left|\frac{\xi_1 + \ldots + \xi_n}{n} - \mathbb{E}\xi\right| > \varepsilon\right) < \delta(\varepsilon, n)$$

где $\delta(\varepsilon, n)$ – некоторая известная функция. Подобные законы говорят, что выражение $\frac{\xi_1(\omega)+\ldots+\xi_n(\omega)}{n}$ отклоняется от матожидания $\mathbb{E}\xi$ на величину большую ε с вероятностью меньше $\delta(\varepsilon, n)$. То есть мы можем гарантировать близость к математическому ожиданию с некоторой вероятностью при достаточно большом количестве измерений. Я не привожу точную формулировку этого утверждения, просто потому что есть куча разных формулировок. Все они основаны на неравенстве Чебышева.

Утверждение (Неравенство Чебышева)**.** *Пусть ξ – случайная величина, такая что $\mathbb{E}\xi = a$ и $\mathbb{D}\xi = \sigma^2 > 0$, тогда для любого числа $t > 0$ верно*

$$P(|\xi - a| \geqslant t\sigma) \leqslant \frac{1}{t^2}$$

Выборочная дисперсия Для оценки дисперсии случайной величины первая формула, которая приходит на ум – это

$$S = S(x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2, \text{ где } \bar{x} = \frac{x_1 + \ldots + x_n}{n}$$

Чтобы проверить несмещенность этой формулы, надо заменить каждое вхождение x_i на независимую $\xi_i \sim \xi$. В этом случае получим, что

$$\mathbb{E}S(\xi_1, \ldots, \xi_n) = \frac{n-1}{n}\mathbb{D}\xi$$

Оказывается, что в этой наивной формуле мы в среднем промахиваемся мимо математического ожидания. Потому вводится понятие несмещенной оценки для выборочной дисперсии

$$S_0 = S_0(x_1, \ldots, x_n) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \text{ где } \bar{x} = \frac{x_1 + \ldots + x_n}{n}$$

С другой стороны, всякие вариации на тему закона больших чисел, показывают, что обе оценки для дисперсии стремятся к самой дисперсии. Я не буду расписывать этот факт более подробно по нескольким причинам. Во-первых, строгие формулировки важны только для математики внутри математики. Во-вторых, от нестрого формулировки надо лишь знать, что в каком-то смысле подобная формула в пределе стремится к желаемой оцениваемой величине. В-третьих, надо понимать, что на практике нам все равно ни холодно ни жарко от разновидностей пределов в теории вероятности, важно лишь, что есть какой-то вид сходимости, а раз он есть, то мы просто надеемся, что наших измерений достаточно, чтобы быть близкими к желаемому пределу (сто процентных гарантий все равно нет).

Векторные измерения Пусть у нас теперь есть некоторый неизвестный случайный вектор $\bar{\xi} = (\xi_1, \ldots, \xi_m)$ на $\mathbb{R}^m$ и пусть у нас есть n независимых измерений $x_1, \ldots, x_n \in \mathbb{R}^m$. Обратите внимание, что мы теперь измеряем вектор, а не число, то есть каждый вектор x_i имеет свои координаты $x_i = (x_{i1}, \ldots, x_{in})$.

Как можно думать про подобные эксперименты. Пусть у вас на черном ящике есть несколько приборов $\xi_1, \ldots, \xi_m$, вообще говоря зависимых между собой. И вы делаете такой эксперимент: снимаете показания с первого прибора, потом со второго и так далее до последнего m-го прибора и записываете все m чисел в вектор x_1. Потом сбрасываете эксперимент в начальное состояние и независимо проводите новую серию таких же измерений и записываете результат в вектор x_2. Потом опять сбрасываете все в начальное состояние и независимо от предыдущих попыток проводите следующий эксперимент и получаете вектор x_3 и т.д. Таким образом внутри i-ой серии, когда вы замеряли координаты вектора x_i, измерения могут быть зависимыми, но между сериями у вас нет зависимостей.

Математическое ожидание векторной величины оценивается так же, как и в скалярном случае

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n} \in \mathbb{R}^m$$

Выборочная матрица ковариации В этом случае самая популярная формула для оценки матрицы ковариации следующая

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^t \in \mathrm{M}_m(\mathbb{R}), \text{ где } \bar{x} = \frac{x_1 + \ldots + x_n}{n} \in \mathbb{R}^m$$

Аналогично дисперсии, если подставить вместо x_i векторы $\bar{\xi}_i$, то окажется, что математическое ожидание от Σ будет промахиваться мимо матрицы ковариации[1]. Потому вводится несмещенная оценка

$$\Sigma_0 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^t \in \mathrm{M}_m(\mathbb{R}), \text{ где } \bar{x} = \frac{x_1 + \ldots + x_n}{n} \in \mathbb{R}^m$$

Оценка корреляции двух случайных величин Пусть есть две случайные величины ξ и η. Напомню, что коэффициентом корреляции между ξ и η называется величина

$$\frac{\mathrm{cov}(\xi, \eta)}{\sqrt{\mathbb{D}\xi}\sqrt{\mathbb{D}\eta}} = \frac{\mathbb{E}(\xi - \mathbb{E}\xi)(\eta - \mathbb{E}\eta)}{\sqrt{\mathbb{D}\xi}\sqrt{\mathbb{D}\eta}}$$

Дисперсии мы уже оценивать умеем, а чтобы оценить ковариацию, надо проделать серию измерений следующего вида: мы измеряем ξ и η одновременно, получаем пару числе (x_1, y_1). Потом независимо измеряем их еще раз и получаем пару чисел (x_2, y_2) и т.д. В результате мы получим набор векторов

$$v_1 = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \ldots, v_n = \begin{pmatrix} x_n \\ y_n \end{pmatrix} \in \mathbb{R}^2$$

Значит мы можем оценить матожидание и матрицу ковариации случайного вектора (ξ, η). Тогда оценка для $\mathrm{cov}(\xi, \eta)$ – будет коэффициент Σ_{12}, а именно, оценочная ковариация равна

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \in \mathbb{R}, \text{ где } \bar{x} = \frac{x_1 + \ldots + x_n}{n} \text{ и } \bar{y} = \frac{y_1 + \ldots + y_n}{n}$$

И несмещенная версия оценочной ковариации выглядит так

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \in \mathbb{R}, \text{ где } \bar{x} = \frac{x_1 + \ldots + x_n}{n} \text{ и } \bar{y} = \frac{y_1 + \ldots + y_n}{n}$$

Метод максимального правдоподобия

Предположим, что у нас есть непрерывная случайная величина ξ, и ее плотность $p(x, \theta_1, \ldots, \theta_k)$ зависит от параметров, которые мы не знаем, но очень хотим узнать. При этом у нас есть n независимых измерений $x_1, \ldots, x_n \in \mathbb{R}$. Вопрос: как найти $\theta_1, \ldots, \theta_k$? Можно воспользоваться следующим соображением. Мы знаем, что измерения пришли из независимых измерений, значит есть такие независимые случайные величины $\xi_1, \ldots, \xi_n \sim \xi$, что $x_1 = \xi_1(\omega), \ldots, x_n = \xi_n(\omega)$. Из-за независимости, у случайного вектора $\bar{\xi} = (\xi_1, \ldots, \xi_n)$ плотность будет задаваться по правилу

$$p_{\bar{\xi}}(z_1, \ldots, z_n, \theta_1, \ldots, \theta_n) = p(z_1, \theta_1, \ldots, \theta_k) \ldots p(z_n, \theta_1, \ldots, \theta_k)$$

Здесь набор $\theta_1, \ldots, \theta_n$ – это параметры нашего распределения, а для каждой ξ_i соответствующая координата в пространстве обозначается z_i. С одной стороны, эта функция задает плотность $\bar{\xi}$, а с другой, нам выпал вектор $(x_1, \ldots, x_n) = \bar{\xi}(\omega)$. Как вы думаете, если нам задана плотность, то какая точка вероятнее всего выпадет в одном эксперименте? Конечно же максимум плотности, просто по ее сельскохозяйственному смыслу – это самая вероятная точка, которая должна выпасть. Значит, нам надо подобрать параметры $\theta_1, \ldots, \theta_k$ так, чтобы

[1] А именно, $\mathbb{E}\Sigma$ будет матрица ковариации умноженная на $\frac{n-1}{n}$.

точка $(x_1, \ldots, x_n)$ оказалась максимумом функции $p_\xi(z, \theta)$. То есть надо решить следующую оптимизационную задачу

$$y = p(z_1, \theta_1, \ldots, \theta_k) \ldots p(z_n, \theta_1, \ldots, \theta_k) \to \max_{\theta_1, \ldots, \theta_k}$$

Заметим, что максимизировать некоторую функцию $\phi(\theta)$, принимающую только положительные значения – это то же самое, что максимизировать функцию $\ln(\phi(\theta))$. Так как ln монотонно возрастает, то на такую процедуру можно смотреть, как на замену координаты y для значения функции, то есть мы просто растягиваем по вертикали график функции $\phi(\theta)$ и не меняем ее точек максимума и минимума, но меняем сами значения в этих точках. Тогда задача оптимизации превращается в следующую

$$\sum_{i=1}^{n} \ln p(z_i, \theta_1, \ldots, \theta_k) \to \max_{\theta_1, \ldots, \theta_k}$$

И часто удобно поменять знак у этого выражения и минимизировать его.

Давайте продемонстрируем, как работает метод на примере гауссова распределения. Пусть ξ имеет распределение с плотностью

$$p(x, a, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-a)^2}{2\sigma^2}}$$

Но мы не знаем конкретных значений для $a \in \mathbb{R}$ и $\sigma > 0$, зато знаем несколько независимых измерений этой случайной величины: $x_1, \ldots, x_n \in \mathbb{R}$. Тогда нам надо максимизировать следующее выражение

$$\phi(a, \sigma) = \sum_{i=1}^{n} \left(-\frac{(x_i - a)^2}{2\sigma^2} - \ln(\sigma) - \ln(\sqrt{2\pi}) \right)$$

Во-первых, мы можем изменить знак у этого выражения и минимизировать полученную функцию. Во-вторых, мы можем отбросить константное слагаемое $\ln(\sqrt{2\pi})$, так как прибавление константы не меняет минимумов функции. Потому будем минимизировать выражение

$$\phi(a, \sigma) = \sum_{i=1}^{n} \left(\frac{(x_i - a)^2}{2\sigma^2} + \ln(\sigma) \right) \to \min_{a, \sigma}$$

Найдем критические точки

$$\frac{\partial \phi}{\partial a} = \sum_{i=1}^{n} -\frac{x_i - a}{\sigma^2} = 0$$
$$\frac{\partial \phi}{\partial \sigma} = \sum_{i=1}^{n} \left(-\frac{(x_i - a)^2}{\sigma^3} + \frac{1}{\sigma} \right) = 0$$

Из первого уравнения получаем выражение для a и, подставив его во второе, получаем выражение для σ. Выражения будут иметь вид

$$a = \frac{x_1 + \ldots + x_n}{n} = \bar{x}$$
$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Полученные оценки и называются оценками максимального правдоподобия для гауссова распределения.

Давайте проверим Гессиан, что действительно получилась точка минимума. Вторые частные производные будут

$$\frac{\partial^2 \phi}{\partial a^2} = \frac{n}{\sigma^2}$$
$$\frac{\partial^2 \phi}{\partial a \partial \sigma} = 2 \sum_{i=1}^{n} \frac{x_i - a}{\sigma^3}$$
$$\frac{\partial^2 \phi}{\partial \sigma^2} = 3 \sum_{i=1}^{n} \frac{(x_i - a)^2}{\sigma^4} - \frac{n}{\sigma^2}$$

Теперь надо вместо a и σ подставить выражения из оценок максимального правдоподобия, потому что мы считаем Гессиан не в произвольной точке, а в той, которую хотим проверить на минимум. Тогда получим

$$\begin{aligned} \frac{\partial^2\phi}{\partial a^2} &= \frac{n}{\sigma^2} \\ \frac{\partial^2\phi}{\partial a\partial\sigma} &= 0 \qquad \text{то есть} \quad H(\phi) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{2n}{\sigma^2} \end{pmatrix} \\ \frac{\partial^2\phi}{\partial\sigma^2} &= \frac{2n}{\sigma^2} \end{aligned}$$

А тут легко видеть, что это положительно определенная матрица (здесь вместо σ^2 стоит выражение из оценки максимального правдоподобия, но для краткости я его оставил в виде σ^2).

PCA и SVD

Пусть у нас задан случайный вектор ξ в пространстве $\mathbb{R}^m$ и мы для него намерели независимо несколько самплов $x_1, \ldots, x_n \in \mathbb{R}^m$. Первым делом мы можем оценить математическое ожидание и матрицу ковариации для ξ по формулам

$$\mathbb{E}\xi = \bar{x} = \frac{x_1 + \ldots + x_n}{n} \quad \Sigma_0 = \frac{1}{n-1}(x_i - \bar{x})(x_i - \bar{x})^t$$

Давайте составим матрицу из всех самплов $X = (x_1|\ldots|x_n)$ и сдвинем каждый сампл на выборочное математическое ожидание $Y = (x_1 - \bar{x}|\ldots|x_n - \bar{x}) = X - (\bar{x}|\ldots|\bar{x})$.

А теперь представьте, что вы почувствовали непреодолимое желание сделать SVD для матрицы Y. Давайте посмотрим, к чему это может привести. На помню, что алгоритм для нахождения SVD для матрицы Y начинается с того, что надо диагонализировать матрицу YY^t. Но если мы применим к ней блочные формулы, то получим, что $YY^t = (x_1 - \bar{x})(x_1 - \bar{x})^t + \ldots + (x_n - \bar{x})(x_n - \bar{x})^t$, что с точностью до коэффициента совпадает с матрицей выборочной ковариации. А значит диагонализация выборочной ковариации – это первый шак в SVD для Y.

Напомню, что под SVD для Y мы понимаем разложение $Y = UDV^t$, где $U \in \mathrm{M}_m(\mathbb{R})$, $V \in \mathrm{M}_n(\mathbb{R})$ и обе ортогональные (то есть $UU^t = E$ и $VV^t = E$) а матрица $D \in \mathrm{M}_{m\,n}(\mathbb{R})$ и на ее главное диагонали стоят числа $\sigma_1 \geqslant \sigma_2 \geqslant \ldots \geqslant \sigma_r > 0$ и быть может нули. В этом случае $YY^t = UDD^tU^t$, где DD^t будет диагональной матрицей размера m на m с числами $\sigma_1^2 \geqslant \ldots \geqslant \sigma_r^2 > 0$ на диагонали и быть может нулями. В этом случае числа $\{\sigma_1^2, \ldots, \sigma_r^2\}$ будут ненулевыми точками спектра YY^t (то есть не нулевые собственные значения этой матрицы, которые совпадают с ненулевыми корнями ее хар многочлена или минимального многочлена). А матрица $U = (u_1|\ldots|u_m)$ будет состоять из собственных векторов для матрицы YY^t причем они все будут ортогональны и длины один относительно стандартного скалярного произведения $(x, y) = x^t y$. При этом u_1 будет собственным для σ_1^2 и так далее до u_r, остальные u_i будут собственными для 0.

На SVD разложение можно посмотреть вот как $D = U^tYV$. То есть мы стартовали с матрицы из несмещенных самплов Y и теперь с помощью матрицы U пытаемся поменять координаты в пространстве $\mathbb{R}^m$, где живут наши самплы, а с помощью матрицы V пытаемся комбинировать наши самплы между собой. Часто операция комбинирования самплов между собой не очень физически осмысленна, потому давайте ее проигнорируем и рассмотрим равенство $U^tY = DV^t$. Тогда вот какой смысл у написанного. Мы стартовали с Y. Потом выбрали для выборочной матрицы ковариации Σ_0 собственные векторы (длины один и взаимногортогональные) и сложили их в матрицу $U = (u_1|\ldots|u_m)$. Далее привели матрицу Σ_0 к главным осям, то есть сделали замену стандартного базиса на базис $u_1, \ldots, u_m$. При этом координаты наших самплов как раз

изменятся по правилу $Y \mapsto U^tY = DV^t$. Давайте внимательно посмотрим на правую часть этого равенства

$$U^tY = DV^t = \begin{pmatrix} \sigma_1 & & & & & & 0 \\ & \ddots & & & & & \vdots \\ & & \sigma_r & & & & 0 \\ & & & 0 & & & 0 \\ & & & & \ddots & & \vdots \\ & & & & & 0 \ \ \dots & 0 \end{pmatrix} \begin{pmatrix} v_{11} & \dots & \dots & v_{n1} \\ \vdots & \dots & \dots & \vdots \\ v_{1m} & \dots & \dots & v_{nm} \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix} =$$

$$= \begin{pmatrix} \sigma_1 & & & & \\ & \ddots & & & \\ & & \sigma_r & & \\ & & & 0 & \\ & & & & \ddots \\ & & & & & 0 \end{pmatrix} \begin{pmatrix} v_{11} & \dots & \dots & v_{n1} \\ \vdots & \dots & \dots & \vdots \\ v_{1m} & \dots & \dots & v_{nm} \end{pmatrix}$$

То есть мы у ортогональной матрицы V^t отрезали верхнюю часть. Так как матрица V была ортогональна, то интуитивно, ее координаты вносят одинаковый вклад в итоговую матрицу. Однако, после этого мы каждую координату домножили на весовой коэффициент σ_i, а последние даже на 0. Таким образом после смены координат для несмещенных самплов $Y \mapsto U^tY$ мы отсортировали координаты сверху вниз по их важности, где σ_i означает вес важности координаты. При этом мы видим, что последние координаты мы вообще можем проигнорировать, ибо они стали нулевыми. Кроме того, мы еще можем проигнорировать координаты с малыми σ_i.

По другому еще можно сказать так, мы нашли, что все наши самплы реально жили в подпространстве $\mathbb{R}^r$ и после поворота пространства, отрезав лишние координаты, мы можем считать, что наши самплы имеют вид

$$\begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{pmatrix} \begin{pmatrix} v_{11} & \dots & \dots & v_{n1} \\ \vdots & \dots & \dots & \vdots \\ v_{1r} & \dots & \dots & v_{nr} \end{pmatrix}$$

А откинув еще и малые σ_i мы можем понизить размерность задачи еще сильнее.

Центральная предельная теорема

Оказывается, что нормальное распределение обладает одним удивительным ассимптотическим свойством. В некотором смысле, если много раз делать один и тот же случайный эксперимент с измерением одной и той же случайной величины, то при правильной нормировке, со временем все самплы будут вести себя так, как будто они взялись из нормального распределения. Это означает, что при большом количестве самплов, нам вообще говоря плевать на распределение, все можно считать нормальным. Звучит слишком уж круто, чтобы быть правдой, тем не менее, я постараюсь придать строгий смысл столь серьезному заявлению. Строгая формулировка этого явления называется Центральной предельной теоремой.[2]

Утверждение (Центральная предельная теорема)**.** *Пусть $\xi_1, \dots, \xi_n, \dots$ – последовательность независимых одинаково распределенных случайных величин с математическим ожиданием μ и дисперсией σ^2. Тогда*

$$\frac{\sum_{k=1}^n (\xi_k - \mu)}{\sigma\sqrt{n}} \xrightarrow{\mathcal{D}} N(0,1)$$

Из закона больших чисел следует, что если мы рассмотрим дробь[3]

$$\frac{\sum_{k=1}^n (\xi_k - \mu)}{n} = \bar{x}_n - \mu$$

[2]На самом деле, это не единственный результат такого рода, но один из первых и уж точно один из важнейших. Есть куча других асимптотических результатов аналогичных ЦПТ и в них фигурируют не только нормальное распределение.

[3]Здесь $\bar{x}_n = \frac{1}{n}\sum_{k=1}^n x_k$ содержит индекс n, чтобы подчеркнуть зависимость от количества самплов.

то она обязательно сойдется к нулю, так как средне арифметическое одинаково распределенных независимых случайных величин сходится к математическому ожиданию. А в центральной предельной теореме мы оцениваем на сколько быстро $\bar{x}$ приближается к μ. Если нормировать не так агрессивной и положить

$$\frac{\sum_{k=1}^{n}(\xi_k - \mu)}{\sigma\sqrt{n}} = \sqrt{n}\frac{\bar{x}_n - \mu}{\sigma}$$

то полученное выражение будет себя вести как нормально распределенная случайная величина. Еще неформально это можно запомнить так, при больших n выполнено

$$\bar{x}_n \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

Таким образом выборочное математическое ожидание будет раскидано вокруг настоящего математического ожидания по нормальному закону.

Если расшифровать ЦПТ и вспомнить третье эквивалентное определение сходимости по распределению, то получается, что

$$P\left(\frac{\sum_{k=1}^{n}(\xi_k - \mu)}{\sigma\sqrt{n}} \in [a,b]\right) \to \int\limits_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx$$

То есть при больших значениях n вероятность для случайной величины

$$\frac{\sum_{k=1}^{n}(\xi_k - \mu)}{\sigma\sqrt{n}}$$

считается по формулам для нормального распределения, в частности вероятность попасть в любое подмножество. В примере выше в качестве такого подмножества был выбран интервал $[a,b]$.

Если дополнительно для случайных величин ξ_k существует $|\xi_k^3| = \rho$ (оно не зависит от k, так как все случайные величины одинаково распределены), то можно оценить скорость сходимости к нормальному распределению. Скорость будем оценивать на функции распределения. Пусть

$$\eta_n = \frac{\sum_{k=1}^{n}(\xi_k - \mu)}{\sigma\sqrt{n}} \quad \text{и} \quad \zeta \sim N(0,1)$$

Тогда

$$|P(\eta_n \leqslant x) - P(\zeta \leqslant x)| \leqslant \frac{\rho}{2\sigma^3\sqrt{n}}$$

Но написанное слева есть ни что иное, как функции распределения

$$|F_{\eta_n}(x) - F_\zeta(x)| \leqslant \frac{\rho}{2\sigma^3\sqrt{n}}$$

Константу $1/2$ можно чуть улучшить в правой части неравенство, которое называется неравенством Берри-Эссеена. Благодаря нему мы не только знаем, что η_n сойдется к нормальному распределению, но и для каждого конкретного интервала оценить на сколько близко соответствующая вероятность подойдет к вероятности для нормального распределения, а именно

$$\left| P\left(\frac{\sum_{k=1}^{n}(\xi_k - \mu)}{\sigma\sqrt{n}} \in [a,b]\right) - \int\limits_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx \right| \leqslant \frac{\rho}{\sigma^3\sqrt{n}}$$