Лекция 8. Проверка гипотез

1. Общие принципы теории проверки гипотез

Предположим, что мы имеем некоторую статистическую выборку X. Под гипотезой мы будем понимать некоторое суждение, которое может быть истинным или ложным. При этом истинность или ложность гипотезы зависит от выборки X. Обозначим эту гипотезу через H_0, которую будем называть основной или нулевой гипотезой. Альтернативной гипотезой называется гипотеза H_1, которая истинна тогда, когда ложна H_0. В простейшем случае речь идет о выборе одной из гипотез H_0 или H_1.

Правило, на основании которого мы будем выбирать и отклонять гипотезы называется статистическим критерием. Повторим, что это правило должно основываться только на выборке.

Через $\mathbb{X}$ обозначим все возможные выборки в рассматриваемом случае. При этом $X \in \mathbb{X}$. Если фиксировать какой-либо статистический критерий, то множество $\mathbb{X}$ разобьется на два не пересекающихся подмножества

$$\mathbb{X} = \mathbb{X}_0 \cup \mathbb{X}_1.$$

Если $X \in \mathbb{X}_0$, то принимается основная гипотеза H_0, если $X \in \mathbb{X}_1$, то основная гипотеза отвергается, а принимается альтернативная гипотеза H_1. Множество $\mathbb{X}_0$ называется критической областью.

Выбор критической области может быть реализован на различных принципах. Согласно общему принципу принятия статистических решений: если в выборке (эксперименте) мы наблюдаем маловероятное с точки зрения гипотезы H_0 событие, тогда гипотеза H_0

должна быть отвергнута, как не согласуемая с данными. В противном случае, гипотеза H_0 не противоречит данным и должна быть принята.

Говоря о «маловероятности» мы имеем в виду, что априорно задано некоторое малое число α (обычно $\alpha \in \{0.001, 0.01, 0.05\}$) и область $\mathbb{X}_1$ должна удовлетворять условию

$$\mathcal{P}(X \in \mathbb{X}_1 | H_0) \leq \alpha.$$

В этом случае говорят, что критерий имеет уровень значимости α.

2. Критерий согласия Колмогорова

Критерий Колмогорова применяется для проверки гипотезы о виде распределения случайной величины. Пусть основная гипотеза состоит в том, что выборка X объема n является результатом реализации некоторой случайной величины ξ. Случайная величина ξ задается функцией распределения $F(x)$. С другой стороны, мы можем по выборке построить эмпирическую функцию распределения $F_n(x)$. Обе функции $F(x)$ и $F_n(x)$ заданы на всей числовой оси и являются ограниченными, поэтому можно определить следующую функцию, зависящую от выборки X

$$D_n(X) = \sup_{-\infty < x < \infty} |F_n(x) - F(x)|.$$

Эта функция называется статистикой Колмогорова.

При больших n, например, при $n > 20$ можно применять следующий критерий, называемый критерием Колмогорова

$$\mathcal{P}(\sqrt{n} D_n \leq \lambda_\alpha | H_0) = \alpha,$$

где λ_α определяется из уравнения

$$K(\lambda_\alpha) = 1 - \alpha,$$

где $K(t) = \sum\limits_{k=-\infty}^{\infty} (-1)^k e^{-2k^2t^2}$ есть известная уже нам функция Колмогорова. Для нахождения λ_α можно использовать известные таблицы функции Колмогорова.

3. Критерий согласия χ^2 Пирсона

Критерий Пирсона используют для проверки гипотезы о законе распределения дискретной случайной величины. Пусть основная гипотеза состоит в том, что случайная величина ξ, принимающая M различных значений $\{x_k\}_{k=1}^M$ с вероятностями $p_1, p_2, \ldots, p_M$. По выборке объема N мы строим количество значений x_k, которые мы обозначим через ν_k.

Построим статистику Пирсона следующим образом

$$X_N^2 = \sum_{k=1}^{M} \frac{(\nu_k - Np_k)^2}{Np_k}.$$

Основная гипотеза H_0 принимается, если выполнено условие

$$X_N^2 \leq \chi^2_{M-1,\alpha},$$

где $\chi^2_{M-1,\alpha}$ находится по таблицам распределения χ^2 с $M-1$ степенью свободы.

4. Примеры использования критерия Пирсона

Пусть мы имеем необычную монету и хотим проверить является ли она честной, т.е. при подбрасывании этой монеты с одинаковой

ли частотой будет орел и решка? Для этого мы подбросили монету 100 раз и получили такие результаты: 55 раз выпал орел и 45 раз выпала решка.

В нашем случае: $M = 2$, $N = 100$, $p_1 = 0.5$, $p_2 = 0.5$, $\nu_1 = 55$, $\nu_2 = 45$. Поэтому значение статистики будет

$$X^2_{100} = \frac{(55 - 100 \cdot 0.5)^2}{100 \cdot 0.5} + \frac{(45 - 100 \cdot 0.5)^2}{100 \cdot 0.5} = 1.$$

При уровне значимости $\alpha = 0.05$ мы принимает основную гипотезу, поскольку

$$1 = X^2_{100} \leq \chi^2_{1,0.05} = 3.8.$$

Рассмотрим пример для оценки эффективности нового лекарства от тяжелой болезни. Пусть по статистики для больных, не получающих новое лекарство, имеет место следующий прогноз:

- 60% болеют в легкой форме;
- 30% болеют в тяжелой форме;
- 60% умирают.

Лекарство испытывалось на 150 больных. При это для этой группы были зафиксированы следующие результаты:

- 105 болело в легкой форме;
- 36% болело в тяжелой форме;
- 9 умерло.

В качестве основной гипотезы мы выберем предположение, что новое лекарство не оказывает статистически значимое влияние на

течение болезни. Таким образом, основная гипотеза — это то, что больные в контрольной группе имеют распределение вероятностей течения болезни

В этом примере число $M = 3$, а $N = 150$, распределение вероятностей $p_1 = 0.6$, $p_2 = 0.3$, $p = 0.1$. Построим статистику Пирсона

$$X^2_{150} = \frac{(105 - 150 \cdot 0.6)^2}{150 \cdot 0.6} + \frac{(36 - 150 \cdot 0.3)^2}{150 \cdot 0.3} + \frac{(9 - 150 \cdot 0.1)^2}{150 \cdot 0.1} = 6.7.$$

При уровне значимости $\alpha = 0.05$ мы получаем

$$6.7 = X^2_{150} > \chi^2_{2,0.05} = 6.$$

Это означает, что основная гипотеза — о бесполезности нового лекарства — должна быть отклонена. Следовательно с вероятностью 0.95 новое лекарство изменяет ход течения болезни.

5. Байесовский подход к проверке гипотез

Центральной идеей байесовского подхода к проверке гипотез является формула Байеса, хорошо известная из теории вероятностей. Напомним эту формулу и понятие условной вероятности. Пусть мы имеем два события A и B. Без ограничения общности, можно считать, что эти события имеют ненулевые вероятности $P(A) > 0$ и $P(B) > 0$. Условной вероятностью события B при условии события A называется величина $P(B|A)$, которая определяется по формуле

$$P(B|A) = \frac{P(AB)}{P(A)},$$

где событие AB есть пересечение событий A и B.

С другой стороны, рассуждая симметрично мы имеем

$$P(A|B) = \frac{P(BA)}{P(A)},$$

но поскольку $AB = BA$, то мы имеем формулу

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}.$$

Эта формула называется формулой Байеса. Значение этой формулы состоит в том, что вычисление условной вероятности $P(A|B)$ можно свести к вычислению условной вероятности $P(B|A)$, что иногда бывает на много проще.

Заметим, что

$$P(AB) = P(B|A)P(A).$$

6. Максимально правдоподобные гипотезы

Пусть теперь мы имеем такой набор событий A_i, $i = 1, \ldots, N$, что

$$\bigcup_{i=1}^{N} A_i = \Omega,$$

где Ω — это множество всех случайных исходов, т.е. множества A_i составляют разбиение множества Ω. Кроме того, мы будем предполагать, что

$$P(A_i) > 0,\ i = 1, 2, \ldots, N$$

и

$$A_i \cap A_j = \varnothing,\ i \neq j.$$

Будем предполагать, что множества A_i являются взаимно исключающими гипотезами, из которых нам необходимо выбрать одну наиболее вероятную. Обычно предполагается, что известны их априорные вероятности $P(A_i)$. В таком случае логично выбрать ту гипотезу, которая имеет наибольшую вероятность. Пусть теперь нам становится известно, что произошло некоторое событие D. В этом

случае нам нужно выбирать ту гипотезу, которая имеет наибольшую условную вероятность

$$P(A_i|D),\ i = 1, 2, \ldots, N.$$

Выбранная таким образом гипотеза называется MAP-гипотеза, где MAP есть аббревиатура «maximum a posteriori», или максимальная апостериорная гипотеза.

Однако во многих случаях вычислить эти вероятности намного сложнее, чем условные вероятности $P(D|A_i)$. Покажем как это можно сделать.

Любое событие D может быть представлено как объединение попарно не пересекающихся событий

$$D = DA_1 \cup DA_2 \cup \cdots \cup DA_N.$$

По формуле полной вероятности мы имеем

$$P(D) = \sum_{i=1}^{N} P(D|A_i)P(A_i),$$

применяя формулу Байеса мы получаем

$$P(A_i|D) = \frac{P(A_i)P(D|A_i)}{\sum\limits_{i=1}^{N} P(D|A_i)P(A_i)}.$$

Заметим, что при выборе максимальной апостериорной гипотезы нам нет необходимости в вычислении знаменателя в последней формуле.

Задача состоит в том, что нам нужно выбрать гипотезу из некоторого набора

$$\{A_1, A_2, \ldots, A_N\}$$

при наличии некоторого события D, которое мы может трактовать как некоторые данные, на основании которых мы будем осуществлять машинное обучение. Разумеется, что событие D может быть очень сложным.

7. Примеры выбора максимально правдоподобных гипотез

В большинстве практических задач событие D, как правило, является является составным из большого количества элементарных событий

$$D = (D_1, D_2, \ldots, D_M),$$

где D_m — это элементарные события, каждое из которых может возникнуть или нет.

В ситуации, когда число M является большим, вычисление события D становится очень сложным, поскольку события D_m могут быть зависимыми событиями.

В практике статистического обучения большую популярность и эффективность получил так называемый наивный байесовский классификатор. Этот подход стоит в том, что мы считаем, что все события $\{D_m\}_{m=1}^{M}$ являются независимыми. В этом случае вероятность $P(D)$ может быть вычислена по формуле

$$P(D) = P(D_1) \cdot P(D_2) \cdot \ \ldots \ \cdot P(D_M).$$

Для выбора гипотез с помощью наивного байесовского классификатора необходимо найти гипотезу A_{i^*}, которая удовлетворяет усло-

вию

$$\max_i P(A_i)P(D_1|A_i)\dots P(D_M|A_i) = \prod_{m=1}^{M} P(A_{i^*})P(D_m|A_{i^*}).$$

Приведем пример использования наивного байесовского классификатора. Пусть в некоторой организации, состоящей из 60% женщин и 40% мужчин мы имеем статистику, которая представлена в следующей таблице.

Таблица VIII.1. Статистика по организации

	$P(D_1)$	$P(D_2)$	$P(D_3)$	$P(D_4)$
женщины	0.3	0.2	0.5	0.15
мужчины	0.6	0.5	0.4	0.3

Здесь:

- D_1 — наличие высшего образования
- D_2 — рост выше 175 см
- D_3 — знание иностранного языка
- D_4 — наличие водительских прав

Пусть некоторый человек из этой группы имеет высшее образование, его рост 172 см, он знает английский язык, не имеет водительских прав. Кто перед нами: мужчина или женщина? Через A_1 обозначим гипотезу, что перед нами женщина, а через A_2 обозначим гипотезу, что перед нами мужчина. Если воспользоваться наивным байесовским классификатором, то можно оценить вероятности следующим образом

$$P(A_1|D) = CP(A_1)P(D_1|A_1)P(D_2|A_1)P(D_3|A_1)P(D_4|A_1) =$$

$$= C \cdot 0.6 \cdot 0.3 \cdot 0.8 \cdot 0.5 \cdot 0.85 = C \cdot 0.0612.$$

и

$$P(A_2|D) = CP(A_2)P(D_1|A_2)P(D_2|A_2)P(D_3|A_2)P(D_4|A_2) =$$

$$= C \cdot 0.4 \cdot 0.6 \cdot 0.5 \cdot 0.4 \cdot 0.7 = C \cdot 0.0336.$$

Здесь через C мы обозначили нормирующую константу, которая для оценки обоих гипотез имеет одинаковое значение и не влияет на выбор гипотезы.

Таким образом, мы видим, что с вероятностью

$$P(A_1|D) = \frac{0.0612}{0.0612 + 0.0336} \approx 0.65$$

перед нами женщина.