Naive Bayes CTR Prediction

Impression attributes

For each impression, and possibly click, we record two attributes:

- time of day: Day or Night
- device: Desktop or Mobile

We have a dataset of this kind:

Announce	Clicked	Time	Device
A	No	Day	Desktop
A	Yes	Day	Mobile
B	No	Night	Desktop
…			

Attributes can be used as *predictors*: if we discover that A is preferred on day time and B on night time, this is a very useful information.

First naïve approach: play 4 different MAB games.

Announce A	**Desktop**	**Mobile**
Day	3 clicks / 100 imps	2 clicks / 50 imps
Night	1 clicks / 30 imps	1 clicks / 80 imps

One MAB game per attribute combination.

Announce B	**Desktop**	**Mobile**
Day	2 clicks / 120 imps	2 clicks / 260 imps
Night	1 clicks / 10 imps	1 clicks / 40 imps

Ad A will be selected more often on day, B on night and so on.

Each time a new user comes, we check the Time and Device attributes, choose the game to play and select the ad to show accordingly.

This approach is natural, but too naïve.

Having 4 MABs means playing each one with 4 times smaller data available.

So, each sample, i.e. each ad's history is less reliable and more varying. We learn much more slowly.

Furthermore, it is not generalizable. If we have 10 attributes, then the possible combinations are 1,024 and we cannot split our dataset in 1,024 histories. We quickly get unable to make reasonable estimation of ads' CTRs.

We want to use 2 predictors to play an unique game instead of 4, or 10 predictors to play an unique game instead of 1,024.

The tool is Bayes theorem.

Remember the formula

$$\Pr(H|D) = \frac{\Pr(D|H) \times \Pr(H)}{\Pr(D|H) \times \Pr(H) + \Pr(D|\neg H) \times \Pr(\neg H)}$$

[The symbol | stays for "conditional on". The symbol ¬ stays for "not"].

Here the hypothesis *H* is "click" and ¬H is "no click".

Data *D* represents the combination of attributes for an event, i.e. an impression and possibly a click.

$$\Pr(H|D) = \frac{\Pr(D|H) \times \Pr(H)}{\Pr(D|H) \times \Pr(H) + \Pr(D|\neg H) \times \Pr(\neg H)}$$

Announce A	**Desktop**	**Mobile**
Day	3 clicks / 100 imps	2 clicks / 50 imps
Night	1 clicks / 30 imps	1 clicks / 80 imps

Let *D* be "Night". The device is not relevant.

We are interested only in choosing between ads A and B using the time now, which is night.

Let us focus ad A.

Pr(H | D) (the *posterior probability)* is the probability that A gets a click when showed at night, i.e. the CTR of A given that time is Night.

We could simply count the historical CTR on A during night. In our example, it is 2 / 110.

Instead, we apply the Bayes Theorem.

$$\Pr(H|D) = \frac{\Pr(D|H) \times \Pr(H)}{\Pr(D|H) \times \Pr(H) + \Pr(D|\neg H) \times \Pr(\neg H)}$$

Announce A	Desktop	Mobile
Day	3 clicks / 100 imps	2 clicks / 50 imps
Night	1 clicks / 30 imps	1 clicks / 80 imps

Pr(D | H) (the *likelihood*) is the probability that it is night when A gets a click. It is 2 / 7.

Pr(H) (the *prior probability*) is the probability that A gets a click independently of the time. It is 7 / 260.

The numerator is (1 / 4) × (7 / 260).

$$\Pr(H|D) = \frac{\Pr(D|H) \times \Pr(H)}{\Pr(D|H) \times \Pr(H) + \Pr(D|\neg H) \times \Pr(\neg H)}$$

Announce A	Desktop	Mobile
Day	3 clicks / 100 imps	2 clicks / 50 imps
Night	1 clicks / 30 imps	1 clicks / 80 imps

The denominator is the explosion of Pr(D) (the marginal probability), i.e. the probability that it is night, independently of success or failure in getting a click.

We can compute it counting nightly impressions over total impressions. It is 110 / 260.

We can also compute each term. Not here, but in complex problems this is necessary.

Try it as an exercise, remembering that $\neg H$ means "the time is Day".

In slide 7, last line, read

the numerator is (1/4) x (7/260)

In slide 8, last line, read

… ¬H means "the ad is not clicked"

$$\Pr(H|D) = \frac{\Pr(D|H) \times \Pr(H)}{\Pr(D|H) \times \Pr(H) + \Pr(D|\neg H) \times \Pr(\neg H)}$$

Announce A	**Desktop**	**Mobile**
Day	3 clicks / 100 imps	2 clicks / 50 imps
Night	1 clicks / 30 imps	1 clicks / 80 imps

Now we can compute the posterior probability: it is

$$\frac{\frac{2}{7} \times \frac{7}{260}}{\frac{110}{260}}$$

The value is 2 / 110, as expected.

We have simply reconstructed the observation with a long computation.

Apparently, we gained nothing. Indeed, in this simple case, we gained nothing.

But, let us move to a more complex problem.

$$\Pr(H|D) = \frac{\Pr(D|H) \times \Pr(H)}{\Pr(D|H) \times \Pr(H) + \Pr(D|\neg H) \times \Pr(\neg H)}$$

Announce A	Desktop	Mobile
Day	3 clicks / 100 imps	2 clicks / 50 imps
Night	1 clicks / 30 imps	1 clicks / 80 imps

The denominator is the explosion of Pr(D) (the marginal probability), i.e. the probability that it is night, independently of success or failure in getting a click.

We can compute it counting nightly impressions over total impressions. It is 110 / 260.

We can also compute each term. Not here, but in complex problems this is necessary.

Try it as an exercise, remembering that $\neg H$ means "the time is Day".

Multiple attributes

Announce A	Desktop	Mobile
Day	3 clicks / 100 imps	2 clicks / 50 imps
Night	0 clicks / 30 imps	1 clicks / 80 imps

⇧ 0 clicks instead of 1

$$P(H|D) = \frac{P(D|H) \times P(H)}{P(D|H) \times P(H) + P(D|\neg H)P(\neg H)}$$

When speaking about the case Night-Desktop, we have

$$P(D|H) \equiv P(Night\ and\ Desktop|Click) = \frac{0}{6} = 0$$

This makes everything vanish and the Bayes formula give 0.

I.e. according to available data, it is impossible to get clicks on desktop by night. Of course, we cannot accept this conclusion.

Announce A	Desktop	Mobile
Day	3 clicks / 100 imps	2 clicks / 50 imps
Night	0 clicks / 0 imps	1 clicks / 80 imps

⇧ 0 clicks and 0 imps

$$P(H|D) = \frac{P(D|H) \times P(H)}{P(D)} = \frac{P(D|H) \times P(H)}{P(D|H) \times P(H) + P(D|\neg H)P(\neg H)}$$

Even worse if we have no examples of a certain case, i.e. no impression.

Not only we have no success (click) but no try, too.

The Bayes formula is not applicable when the dataset contains "holes".

Announce A	Desktop	Mobile
Day	3 clicks / 100 imps	2 clicks / 50 imps
Night	0 clicks / 0 imps	1 clicks / 80 imps

⇧ 0 clicks and 0 imps

$$P(H|D) = \frac{P(D|H) \times P(H)}{P(D)} = \frac{P(D|H) \times P(H)}{P(D|H) \times P(H) + P(D|\neg H)P(\neg H)}$$

Unfortunately, if we have a lot of attributes then possible combined cases are a huge number and some cases will be without examples. As we said, 10 binary attributes split the dataset in 1,024 cases.

The problem is not only with properly empty cases (cells in the data table). If we have 2 impressions and 0 clicks, really we have no significant data. *We need significant samples in every case.*

Announce A	Desktop	Mobile
Day	3 clicks / 100 imps	2 clicks / 50 imps
Night	0 clicks / 0 imps	1 clicks / 80 imps

⇧ 0 clicks and 0 imps

$$P(H|D) = \frac{P(D|H) \times P(H)}{P(D)} = \frac{P(D|H) \times P(H)}{P(D|H) \times P(H) + P(D|\neg H)P(\neg H)}$$

The idea we are going to apply is analyzing data related to single attributes, not to their combinations.

If we are able to estimate the impact of *Night* alone on the CTR and alike for *Desktop* alone, then we can estimate the CTR for *Night and Desktop* recombining the two.

This is key insight for many Data Science applications.

Announce A	Desktop	Mobile
Day	3 clicks / 100 imps	2 clicks / 50 imps
Night	0 clicks / 0 imps	1 clicks / 80 imps

⇧ 0 clicks and 0 imps

$$P(H|D) = \frac{P(D|H) \times P(H)}{P(D)} = \frac{P(D|H) \times P(H)}{P(D|H) \times P(H) + P(D|\neg H)P(\neg H)}$$

Let us focus the case where *D* is *Night and Desktop*.

Then P(D|H) is P(Night and Desktop | Click).

The history says it is 0 because in the past we never observed a user to click after seeing ad A on a desktop by night.

This does not mean it is impossible in the future.

Announce A	Desktop	Mobile
Day	3 clicks / 100 imps	2 clicks / 50 imps
Night	0 clicks / 0 imps	1 clicks / 80 imps

↑ 0 clicks and 0 imps

$$P(H|D) = \frac{P(D|H) \times P(H)}{P(D)} = \frac{P(D|H) \times P(H)}{P(D|H) \times P(H) + P(D|\neg H)P(\neg H)}$$

Here P(D|H) is $P(Night\ and\ Desktop|Click)$.

We estimate this probability with this formula $P(Night|Click) \times P(Desktop|Click)$

In general the two probabilities are **not** equal.

In our example

$$P(Night\ and\ Desktop|Click) = 0\ (historical\ frequence)$$

$$P(Night|Click) \times P(Desktop|Click) = \frac{1}{6} \times \frac{3}{6}$$

Announce A	Desktop	Mobile
Day	3 clicks / 100 imps	2 clicks / 50 imps
Night	0 clicks / 0 imps	1 clicks / 80 imps

↑ 0 clicks and 0 imps

$$P(H|D) = \frac{P(D|H) \times P(H)}{P(D)} = \frac{P(D|H) \times P(H)}{P(D|H) \times P(H) + P(D|\neg H)P(\neg H)}$$

We have reconstructed the cell Night and Desktop multiplying the impact of the row Night by the impact of the column Desktop.

This is correct only if the two attributes Time and Device are independent,

i.e. only if knowledge about one attribute does not help us to predict the other.

In our example this does not hold: if we know now it is day, we predict that the device getting an impression is more likely to be Desktop. Vice versa if we know now it is night.

The attributes are not independent. We are doing an incorrect thing.

Announce A	Desktop	Mobile
Day	3 clicks / 100 imps	2 clicks / 50 imps
Night	0 clicks / 0 imps	1 clicks / 80 imps

⬆ 0 clicks and 0 imps

$$P(H|D) = \frac{P(D|H) \times P(H)}{P(D)} = \frac{P(D|H) \times P(H)}{P(D|H) \times P(H) + P(D|\neg H)P(\neg H)}$$

Well, we do it even knowing it is a mistake!

We pretend the attributes are independent and compute probabilities of combined events multiplying probabilities of elementary events.

This enables us to use the Bayes formula even when a cell is empty. In general, we use fictitious data with samples bigger than in reality.

The resulting predictions are biased.

In practice, often the benefits exceed the cost of making biased predictions. Predictors of this kind are largely used, often with good results.

Announce A	Desktop	Mobile
Day	3 clicks / 100 imps	2 clicks / 50 imps
Night	0 clicks / 0 imps	1 clicks / 80 imps

⇧ 0 clicks and 0 imps

$$P(H|D) = \frac{P(D|H) \times P(H)}{P(D)} = \frac{P(D|H) \times P(H)}{P(D|H) \times P(H) + P(D|\neg H)P(\neg H)}$$

$$P(D|H) = P(Night|Click) \times P(Desktop|Click) = 1/6 \times 3/6$$
$$P(H) = P(Click) = 6/230$$
$$P(D|\neg H) = P(Night|\neg Clik) \times P(Desktop|\neg Click) = 79/224 \times 97/224$$
$$P(\neg H) = P(\neg Click) = 224/230$$
$$P(H|D) = P(Click|Night\ and\ Desktop)$$
$$= \frac{(\frac{1}{6} \times \frac{3}{6}) \times \frac{6}{230}}{[\left(\frac{1}{6} \times \frac{3}{6}\right) \times \frac{6}{230}] + [(\left(\frac{79}{224}\right) \times \frac{97}{224}) \times \frac{224}{230}]}$$

=1/69 approximately. Small, not zero!

This is an estimated probability, not an empirical frequency.

Factors were empirical frequencies, i.e. counts.

This is called *Naïve Bayes* method in the sense it assumes something false.

Indeed, it is not innocent: we do so because we hope to pay a small price fro the benefit of having a bayesian predictor.

Normally, naïve bayesian predictors work well when you are interested not in relative probabilities (A is twice as likely than B) but in ranks (A is more probable than B).

In multi armed bandit problems this is often good enough.

An interesting option is to use a Naïve Bayes predictor inside a Thompson Sampling algorithm.