Coin tosses

Daniel Hsu (COMS 4771)

Binary predictions

A coin is tossed, and your goal is to predict the outcome (which is either "heads" or "tails"). How should you predict?

If you know the initial conditions of the coin toss, as well as a lot of physics, then (in a non-quantum universe) you can put this knowledge to use to determine exactly how the coin will land. But suppose you don't have all of this knowledge.

If the coin is "fair", then intuitively it doesn't matter how we predict. But if the coin is "biased", then predicting one way may be better than the other.

We'll use a *statistical model* of the problem to motivate a prediction strategy, as well as to evaluate the quality of various strategies. In this model, the outcome of the coin toss is random; it is "heads" with some probability, say, p; and it is "tails" with the remaining probability $1 - p$. We'll encode "heads" by 1, and "tails" by 0, so the outcome is a *random variable* Y. The number p is a *parameter* of the model; the possible values it can taken on, namely the interval $[0, 1]$, is the *parameter space*. This particular model is the family of *Bernoulli distributions* $\{\mathrm{Bern}(p) : p \in [0, 1]\}$; we say that the distribution of Y is $\mathrm{Bern}(p)$ by writing

$$Y \sim \mathrm{Bern}(p).$$

If you know the parameter p, then how should you predict? Here is one strategy:

- If $p > 1/2$, then predict "heads".
- If $p < 1/2$, then predict "tails".
- If $p = 1/2$, doesn't matter. But, for concreteness, predict "tails".

Using this strategy, what is the probability that you predict incorrectly? A simple calculation shows that it is $\min\{p, 1 - p\}$.

Can any strategy have a smaller probability of predicting incorrectly? No. For example, if $p > 1/2$ and you predict "tails", then your probability of predicting incorrectly is at least p, which is more than $1 - p$.

Of course, this all assumes you know the parameter p exactly. In the next section, we'll discuss what can be done when you don't know p.

The plug-in principle and the iid model

In many prediction problems where a statistical model may be used, such as in the problem described above, the parameters of the model are generally not exactly known. In the case above, the parameter p is needed to determine the optimal prediction. Without knowledge of p, we need another way to derive a prediction.

If the outcomes of n previous tosses of the given coin are available, and the goal is to make a prediction as in the preceding problems for an $(n + 1)$-th toss, then we may again use a statistical model to derive a good prediction. In this model, the outcomes $Y_1, \ldots, Y_n, Y$ of the $n + 1$ tosses are *independent and identically*

distributed (*iid*); the first n outcomes $Y_1, \ldots, Y_n$ are observed, and the $(n+1)$-th Y is the outcome to predict. We write this as

$$Y_1, \ldots, Y_n, Y \sim_{\text{iid}} P,$$

where P is the (unknown) distribution of Y. We think of $(Y_i)_{i=1}^n$ as *data* that can be used to derive a prediction of Y. The optimal prediction for the outcome Y is given as some formula depending on unknown aspects of the distribution of Y. The *plug-in principle* prescribes the following steps to form a prediction in this *iid model*:

1. Estimate the unknowns based on the observed outcomes.
2. Plug-in these estimates into the formula for the optimal prediction.

The iid model is ubiquitous in machine learning, as it provides a very simple connection between the observed data and the target of prediction.

Using maximum likelihood estimation

When the statistical model for our problem is a *parametric model*, there is a well-weathered method for deriving estimators based on the *maximum likelihood principle*. Recall that a *parametric* model is a family of probability distributions $\{P_\theta : \theta \in \Theta\}$ for a random variable Z, where the family is indexed by a set Θ. The set Θ is called the *parameter space*, and $\theta \in \Theta$ is the parameter of the distribution P_θ. In many cases, the random variable Z may actually be a vector of several random variables; in such cases, we say Z is a random vector. Similarly, in many cases, each $\theta \in \Theta$ is actually a vector of multiple parameters, so we call each such θ a parameter vector. The *likelihood* $\mathcal{L}(\theta)$ of a parameter vector $\theta \in \Theta$ given an observation $Z = z$ is the probability of $Z = z$ under the distribution P_θ. The *maximum likelihood estimator* (MLE) is

$$\hat{\theta} := \arg\max_{\theta \in \Theta} \mathcal{L}(\theta)$$

i.e., the parameter vector with the highest likelihood.[1] Note that by the strict monotonicity of the logarithm function, the MLE is also the maximizer of the *log-likelihood* function $\ln \mathcal{L}$.

We return to the problem of predicting of the outcome of a coin toss, where the outcome must be either 1 ("heads") or 0 ("tails"). There, we have

$$Y_1, \ldots, Y_n, Y \sim_{\text{iid}} \text{Bern}(p)$$

for some unknown parameter $p \in [0, 1]$. The MLE for p given $(Y_1, \ldots, Y_n) = (y_1, \ldots, y_n)$ is the maximizer of the log-likelihood

$$\begin{aligned} \ln \mathcal{L}(p) &= \ln \prod_{i=1}^n p^{y_i} (1-p)^{1-y_i} \\ &= \sum_{i=1}^n y_i \ln p + (1 - y_i) \ln(1-p). \end{aligned}$$

We analytically determine the maximizer using calculus. First, we find the critical points of $\ln \mathcal{L}$ (i.e., zeros of the derivative of $\ln \mathcal{L}$). The derivative of $\ln \mathcal{L}$ is

$$\frac{1}{p} \sum_{i=1}^n y_i - \frac{1}{1-p} \sum_{i=1}^n (1 - y_i),$$

and it is equal to zero exactly when

$$p = \frac{1}{n} \sum_{i=1}^n y_i.$$

[1] In general, there may be multiple distinct θ with the same highest likelihood, i.e., the MLE is not unique. And in other cases, there may not be any single θ with highest likelihood, i.e., the MLE does not exist.

Next, we determine whether the critical point is a maximizer, a minimizer, or a saddle point. The second-derivative of $\ln \mathcal{L}$ is

$$-\frac{1}{p^2}\sum_{i=1}^{n} y_i - \frac{1}{(1-p)^2}\sum_{i=1}^{n}(1-y_i),$$

which is always non-positive (for $0 < p < 1$).[2] Hence the critical point is a maximizer. Thus, the MLE for p given $(Y_1, \ldots, Y_n) = (y_1, \ldots, y_n)$ is

$$\hat{p} = \hat{p}(y_1, \ldots, y_n) := \frac{1}{n}\sum_{i=1}^{n} y_i.$$

This estimator is used via the plug-in principle to derive the prediction strategy

$$\hat{y}(y_1, \ldots, y_n) := \mathbb{1}\{\hat{p}(y_1, \ldots, y_n) > 1/2\}.$$

Probability of an error

In the iid model, the probability that $\hat{Y} := \hat{y}(Y_1, \ldots, Y_n)$ does not correctly predict Y can be expressed as

$$\begin{aligned}\mathbb{P}(\hat{Y} \neq Y) &= \mathbb{P}(\hat{y}(Y_1, \ldots, Y_n) \neq Y) \\ &= \mathbb{P}(Y_1 + \cdots + Y_n > n/2) \cdot \mathbb{P}(Y = 0) + \mathbb{P}(Y_1 + \cdots + Y_n \leq n/2) \cdot \mathbb{P}(Y = 1).\end{aligned}$$

Suppose $Y \sim \text{Bern}(p)$ for some $p > 1/2$. Using a tail bound for sums of iid Bernoulli random variables (given in the next section), this probability can be bounded as

$$\begin{aligned}\mathbb{P}(\hat{Y} \neq Y) &= (1-p) + (2p-1) \cdot \mathbb{P}(Y_1 + \cdots + Y_n \leq n/2) \\ &\leq (1-p) + (2p-1) \cdot e^{-n \cdot \text{RE}(1/2, p)},\end{aligned}$$

where

$$\text{RE}(a, b) := a \ln \frac{a}{b} + (1-a) \ln \frac{1-a}{1-b}$$

is the *relative entropy* between $\text{Bern}(a)$ and $\text{Bern}(b)$. If instead $Y \sim \text{Bern}(p)$ for some $p \leq 1/2$, the probability of a prediction error is

$$\begin{aligned}\mathbb{P}(\hat{Y} \neq Y) &= p + (1-2p) \cdot \mathbb{P}(Y_1 + \cdots + Y_n > n/2) \\ &\leq p + (1-2p) \cdot e^{-n \cdot \text{RE}(1/2, p)}.\end{aligned}$$

Hence, in either case,

$$\mathbb{P}(\hat{Y} \neq Y) \leq \min\{p, 1-p\} + |2p-1| \cdot e^{-n \cdot \text{RE}(1/2, p)}.$$

Recall that the optimal prediction predicts incorrectly with probability $\min\{p, 1-p\}$. The relative entropy is always non-negative, and $\text{RE}(a, b) = 0$ if and only if $a = b$. Therefore, the probability from the above displayed equation exceeds this by a quantity that goes to zero with n exponentially fast.

Probability tail bounds

Theorem (Tail bounds for sums of iid Bernoulli random variables). Let $X_1, \ldots, X_n \sim_{\text{iid}} \text{Bern}(p)$, and let $S := X_1 + \cdots + X_n$. For any $0 \leq \ell \leq p \leq u \leq 1$,

$$\begin{aligned}\mathbb{P}(S \leq n \cdot \ell) &\leq e^{-n \cdot \text{RE}(\ell, p)}, \\ \mathbb{P}(S \geq n \cdot u) &\leq e^{-n \cdot \text{RE}(u, p)}.\end{aligned}$$

[2] The cases $p = 0$ and $p = 1$ need to be treated differently.

Proof. We just show the first inequality; the second one is proved similarly. We can also assume that $0 < \ell < p < 1$, since the other cases are trivial.

Let f_p denote the probability mass function (pmf) for $(X_1, \ldots, X_n)$. Let $E \subseteq \{0,1\}^n$ be the set of outcomes $x = (x_1, \ldots, x_n)$ where $\sum_{i=1}^n x_i \leq n \cdot \ell$. Then $\mathbb{P}(S \leq n \cdot \ell) = \sum_{x \in E} f_p(x)$. Our goal is to bound this latter sum by $e^{-n \cdot \mathrm{RE}(\ell, p)}$.

Let f_ℓ denote the pmf for $(X'_1, \ldots, X'_n)$ where $X'_1, \ldots, X'_n \sim_{\text{iid}} \mathrm{Bern}(\ell)$. The proof proceeds by comparing f_p to f_ℓ at every $x \in E$. Indeed, fix any $x \in E$, and let k_x be the number of i's such that $x_i = 1$. Since $x \in E$, we have $k_x \leq n \cdot \ell$. Observe that because $p/\ell > 1$, we have $(p/\ell)^{k_x} \leq (p/\ell)^{n \cdot \ell}$; similarly, because $(1-p)/(1-\ell) < 1$, we have $((1-p)/(1-\ell))^{n-k_x} \leq ((1-p)/(1-\ell))^{n \cdot (1-\ell)}$. Therefore

$$\frac{f_p(x)}{f_\ell(x)} = \frac{p^{k_x}(1-p)^{n-k_x}}{\ell^{k_x}(1-\ell)^{n-k_x}} = \left(\frac{p}{\ell}\right)^{k_x} \left(\frac{1-p}{1-\ell}\right)^{n-k_x} \leq \left(\frac{p}{\ell}\right)^{n \cdot \ell} \left(\frac{1-p}{1-\ell}\right)^{n \cdot (1-\ell)}.$$

Because the above inequality holds for every $x \in E$ and $\sum_{x \in E} f_\ell(x) \leq 1$,

$$\sum_{x \in E} f_p(x) \leq \sum_{x \in E} f_\ell(x) \left(\frac{p}{\ell}\right)^{n \cdot \ell} \left(\frac{1-p}{1-\ell}\right)^{n \cdot (1-\ell)} \leq \left(\frac{p}{\ell}\right)^{n \cdot \ell} \left(\frac{1-p}{1-\ell}\right)^{n \cdot (1-\ell)} = e^{-n \cdot \mathrm{RE}(\ell, p)}.$$

∎

The distribution of the random variable S from the tail bound above is called the *binomial distribution* with n trials and success probability p, written $S \sim \mathrm{Bin}(n, p)$. The expected value of S is $n \cdot p$, and the tail bound says that that chance that S deviates from this number by more than a constant factor of n is exponentially small in n.

Although it is extremely unlikely for $S \sim \mathrm{Bin}(n, p)$ to deviate from its mean by magnitudes proportional to n, one can expect deviations of a smaller size. The variance of S gives a bound on the expected magnitude of the deviation:

$$\mathbb{E}[|S - \mathbb{E}(S)|] \leq \sqrt{\mathbb{E}[(S - \mathbb{E}(S))^2]} = \sqrt{\mathrm{var}(S)} = \sqrt{n \cdot p(1-p)}.$$

The inequality in the first step follows from *Jensen's inequality*.