Estimating Probabilities from data

Remember that the Bayes Optimal classifier: "all we needed was" $P(Y|X)$. Most of supervised learning can be viewed as estimating $P(X, Y)$.

There are two cases of supervised learning:

- When we estimate $P(Y|X)$ directly, then we call it *discriminative learning.*
- When we estimate $P(X|Y)P(Y)$, then we call it *generative learning.*

Some machine learning algorithms (e.g. kNN) do not estimate $P(Y|X)$ but only the function $f(x) = \text{argmax}_y P(Y = y|x)$. This is also considered discriminative learning. Not estimating the probabilities can provide some flexibility in terms of what approach is used, but one loses the advantage of knowing probability estimates of the labels and may not have a good reading on the certainty of a prediction.

So, how can we estimate probabilities from data?
There are many ways to estimate probabilities from data.

Simple scenario: coin toss

Suppose you find a coin and it's ancient and very valuable. **Naturally**, you ask yourself, "What is the probability that it comes up heads when I toss it?" You toss it $n = 10$ times and get results: $H, T, T, H, H, H, T, T, T, T$. What is $P(H)$?

We observed n_H heads and n_T tails. So, intuitively,

$$P(H) \approx \frac{n_H}{n_H + n_T} = 0.4$$

Can we derive this formally?

Maximum Likelihood Estimation (MLE)

Let $P(H) = \theta$. θ, however, is unknown and all we have is D (sequence of heads and tails). So, what we can do to estimate θ is to choose its value such that the data is most likely.

MLE Principle: Find $\hat{\theta}$ to maximize the likelihood of the data, $P(D \mid \theta)$:

$$\hat{\theta}_{MLE} = argmax_\theta \, P(D \mid \theta)$$

For the sequence of coin flips we can use the **binomial distribution** to model $P(D \mid \theta)$:

$$P(D \mid \theta) = \binom{n_H + n_T}{n_H} \theta^{n_H} (1 - \theta)^{n_T}$$

Now,

$$\begin{aligned}
\hat{\theta}_{MLE} &= argmax_\theta \binom{n_H + n_T}{n_H} \theta^{n_H} (1 - \theta)^{n_T} \\
&= argmax_\theta \, log \binom{n_H + n_T}{n_H} + n_H \cdot log(\theta) + n_T \cdot log(1 - \theta) \\
&= argmax_\theta \, n_H \cdot log(\theta) + n_T \cdot log(1 - \theta)
\end{aligned}$$

We can now solve for θ by taking the derivative and equating it to zero. This results in

$$\frac{n_H}{\theta} = \frac{n_T}{1 - \theta} \implies n_H - n_H\theta = n_T\theta \implies \theta = \frac{n_H}{n_H + n_T}$$