CMPT 210: Probability and Computation

Lecture 19

Sharan Vaswani

July 19, 2022

Assignment 3 late submission today

Warmup – Tossing coins

Q: We have a coin such that Pr[heads] $= p$. We toss this coin 5 times independently and record the observations. What is the probability that we get 2H and 3T in the 5 tosses.

If X is the r.v. that is equal to the number of heads in 5 tosses, then $X \sim \text{Bin}(5, p)$ and hence, Pr[3H and 2T] $= \binom{5}{2} p^2 (1-p)^3$.

Warmup – Tossing coins

Q: We have a coin such that Pr[heads] $= p$. We toss this coin 5 times independently. What is the probability that we get the following sequence of observations – HTHHT.

$$\begin{aligned}\Pr[\text{observe HTHHT}] &= \Pr[\text{Toss 1 is H} \cap \text{Toss 2 is T} \cap \ldots \cap \text{Toss 5 is T}] \\ &= \Pr[\text{Toss 1 is H}]\ \Pr[\text{Toss 2 is T} \ldots \Pr[\text{Toss 5 is H}] \\ &\qquad\qquad \text{(Since the tosses are independent)} \\ \Pr[\text{observe HTHHT}] &= p^3(1-p)^2\end{aligned}$$

Q: If I use a different coin that has Pr[heads] $= q$ and repeat the same experiment, what is the probability that we get the following sequence of observations – HTHHT?

By the same reasoning as before, Pr[observe HTHHT] $= q^3(1-q)^2$.

Hence, we can say that Pr[observe HTHHT|coin has Pr[heads $= p$]] $= p^3(1-p)^2$.

Questions?

Estimating the bias of a coin

Let us "invert" this reasoning – suppose we took a coin and want to estimate its bias i.e. figure out what is the Pr[heads].

To do this, we take the coin and perform an experiment – toss the coin 5 times and record the observations. Suppose we get $HTHHT$ as the sequence of observations.

This sequence of observations that we got is referred to as the **data** and denoted by $\mathcal{D}$. Using $\mathcal{D}$, we wish to **estimate** the bias of the coin.

If we "guess" the bias of the coin to be p, then the probability that we would see $\mathcal{D}$ is equal to $p^3(1-p)^2$ (by exactly the same reasoning as before). Formally,

$$\Pr[\mathcal{D}|p] = p^3(1-p)^2$$

This is referred to as the **likelihood** of seeing the data (given p).

Estimating the bias of a coin

The standard way to "fit" the data is **maximum likelihood estimation**. For this, the standard procedure is to compute $\hat{p}$ that maximizes the likelihood of observing $\mathcal{D}$.

Formally,

$$\hat{p} = \arg\max_{p} \Pr[\mathcal{D}|p]$$

Here, $\arg\max_p$ returns the value of p that maximizes the likelihood. $\hat{p}$ is the **statistical estimate** of p (similar to what we saw in the Voter Poll example) and is also referred to as the **maximum likelihood estimator (MLE)**.

It is equivalent and more convenient to calculate the minimizer of the **negative log-likelihood** (NLL) (since log is a monotonic function). The NLL is also referred to as the **loss function**. Formally,

$$\hat{p} = \arg\min_{p} -\log(\Pr[\mathcal{D}|p])$$

Estimating the bias of a coin

Let us compute the MLE for the bias of the coin. Recall that $\Pr[\mathcal{D}|p] = p^3(1-p)^2$.

$$-\log(\Pr(\mathcal{D}|p)) = -3\log(p) - 2\log(1-p) \implies \hat{p} = \arg\min_p[-3\log(p) - 2\log(1-p)]$$

Taking derivatives and setting it to zero,

$$\frac{d[-3\log(p) - 2\log(1-p)]}{dp} = 0 \implies -\frac{3}{\hat{p}} + \frac{2}{1-\hat{p}} = 0 \implies 5\hat{p} = 3 \implies \hat{p} = \frac{3}{5} = 0.6.$$

Checking that this is the minimum by computing the second derivative,

$$\frac{d^2[-3\log(p) - 2\log(1-p)]}{dp^2} = \frac{d[\frac{-3}{p} + \frac{2}{1-p}]}{dp} = +\frac{3}{p^2} + \frac{2}{(1-p)^2} > 0 \qquad \text{(for } p \in (0,1)\text{)}$$

Hence, $\hat{p}$ is the minimum of the NLL.

For this simple example of estimating the bias of a coin, the MLE (for estimating the Pr[heads]) is equal to the average number of heads we saw in $\mathcal{D}$.

Estimating the bias of a coin

Q: Based on the results of our experiment, what should be our "guess"/"prediction" that we get a heads in the next toss of the coin?

We have estimated the bias of the coin to be equal to 0.6. Hence, given the results of our experiment, we should **predict** that we will get a heads with probability 0.6 when we toss this coin again in the future.

We just solved a machine learning problem!

Machine learning

The basic framework in machine learning is to:

- Collect (training) data from the world (in this case, by tossing the coin).
- Construct a model that can explain the observations (in this case, our model was that each toss is independent and follows the same Bernoulli distribution).
- Use the model and $\mathcal{D}$ to construct the likelihood function (in this case, $p^3(1-p)^2$).
- Compute the MLE by minimizing the negative log-likelihood. This is an optimization problem (in this case, it was just taking derivatives) and is referred to **training** the model (in this case, finding the **parameter** $\hat{p}$).
- Use the trained model to make predictions about the future (in this case, predict the probability that the next toss comes up heads). This is referred to as **prediction** or **inference**.

Estimating the bias of a coin

Q: Suppose someone hands a new coin and asks us the following question: I tossed this coin 10 times, I got $4H$ and $6T$. What is the probability that I will see $4H$ and $6T$ in the next 10 tosses of the coin.

In this case, $\mathcal{D}$ (referred to as the **training data**) consists of the $4H$ and $6T$ observations in the 10 tosses of the coin. For computing the MLE for the bias of the coin, recall that it is equal to the average number of heads we got in the 10 tosses. And hence in this case, $\hat{p} = 0.4$.

The question is that of predicting the probability of getting $4H$ and $6T$ in the next 10 tosses (referred to as the **test data**). If X is the r.v. equal to the number of heads in the next 10 tosses of the coin, then given $\mathcal{D}$, $X \sim \text{Bin}(10, 0.4)$.

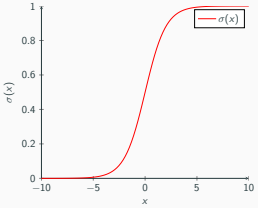
Hence, $\Pr[\text{4H, 6T in the next 10 tosses}|\mathcal{D}] = \binom{10}{4}(0.4)^4(0.6)^6$.

Questions?

Sigmoid function

To extend this concept to more complicated problems, let us introduce the **sigmoid** function.

The **sigmoid** function is defined as: $\sigma : \mathbb{R} \to [0, 1]$: $\sigma(x) := \frac{1}{1+\exp(-x)}$.



Since the range of σ is $[0, 1]$, we will use it to output probabilities. Define parameter $\theta \in \mathbb{R}$ s.t. $p = \sigma(\theta) = \frac{1}{1+\exp(-\theta)}$.

$$1 - p = 1 - \frac{1}{1 + \exp(-\theta)} = \frac{\exp(-\theta)}{1 + \exp(-\theta)} = \frac{1}{1 + \exp(\theta)}$$

Hence, if $p = \sigma(\theta)$, $1 - p = \sigma(-\theta)$.

Sigmoid function

σ is an invertible function and hence there is a one-one mapping from θ to p (every p can be specified by specifying the equivalent θ). Formally, since $p = \sigma(\theta)$ and $1 - p = \sigma(-\theta)$.

$$\frac{p}{1-p} = \frac{\sigma(\theta)}{\sigma(-\theta)} = \frac{1+\exp(\theta)}{1+\exp(-\theta)} = \frac{\exp(\theta)\ (1+\exp(\theta))}{1+\exp(\theta)} = \exp(\theta) \implies \log\left(\frac{p}{1-p}\right) = \theta$$

Recall from Assignment 2 that $\frac{p}{1-p}$ is referred to as the **odds**. Hence the sigmoid transformation is equivalent to choosing the parameter θ to represent the **log-odds**.

Back to estimating the bias of a coin

Recall that when $\mathcal{D} = HTHHT$, $\Pr[\mathcal{D}|p] = p^3(1-p)^3$. Since there is a one-one mapping from p to θ, we can represent the likelihood in terms of θ. Formally,

$$\Pr[\mathcal{D}|\theta] = [\sigma(\theta)]^3[\sigma(-\theta)]^2 = \frac{1}{(1+\exp(-\theta))^3\ (1+\exp(\theta))^2} = \frac{(\exp(\theta))^3}{(1+\exp(\theta))^5}$$

Let us write down the NLL and minimize it w.r.t θ (exactly as we did w.r.t p).

$$-\log(\Pr[\mathcal{D}|\theta]) = -3\log(\exp(\theta)) + 5\ \log(1+\exp(\theta))$$

Computing the derivative and setting it zero,

$$\frac{d[-\log(\Pr[\mathcal{D}|\theta])]}{d\theta} = -3 + \frac{5\exp(\hat{\theta})}{1+\exp(\hat{\theta})} = 0 \implies \frac{\exp(\hat{\theta})}{1+\exp(\hat{\theta})} = \frac{3}{5} \implies \exp(\theta) = \frac{3}{2} \implies \theta = \ln(3/2)$$

Q: Sanity check: What is p when $\theta = \ln(3/2)$?

Ans: $p = \frac{1}{1+\exp(-\theta)} = \frac{\exp(\theta)}{1+\exp(\theta)} = \frac{3/2}{1+3/2} = 3/5 = 0.6$ which is exactly what we had earlier.

Questions?

Estimating the bias of multiple coins

Q: Suppose now we toss 5 different coins and obtain the sequence $HTHHT$. We want to estimate the bias of each of these coins, but have some additional information that the bias of the coin i depends on its weight $x_i \in \mathbb{R}$ (which is known).

We will assume a **linear model** meaning that our model for the bias of coin is:

$$p_i = \sigma(\theta x_i) = \frac{1}{1 + \exp(-\theta x_i)}$$

Here, θ is the **parameter** of our model, x_i (the known weights) for the coins are referred to as the **features**. The model is **linear** because the argument to the sigmoid function is linear in θ.

As before, we need to obtain the MLE $\hat{\theta}$. Writing down the likelihood in terms of θ,

$$\Pr[\mathcal{D}|\{x_1, x_2, \ldots, x_5\}, \theta] = [\sigma(\theta x_1)]\,[\sigma(-\theta x_2)]\,[\sigma(\theta x_3)][\sigma(\theta x_4)][\sigma(-\theta x_5)]$$

Note that now identity of the coins matter because of their weight. Meaning that in the likelihood term, coin 1 with weight x_1 coming up heads is NOT the same as coin 2 with weight x_2 coming up heads.

Estimating the bias of multiple coins

To represent the likelihood in a more compact way, let us define $y_i \in \mathbb{R}$ such that $y_i = 1$ if coin i in $\mathcal{D}$ is a heads and $y_i = -1$ if coin i in $\mathcal{D}$ is a tails. For $\mathcal{D} = HTHHT$, $y_1 = 1$, $y_2 = -1$ and so on. For example i, y_i is referred to as the **label**, hence each toss i is described by the (x_i, y_i) pair referred to as the **input-output** pair or the **feature-label** pair.

$$\Pr[\mathcal{D}|\{x_1, x_2, \ldots, x_5\}, \theta] = [\sigma(y_1\theta x_1)]\,[\sigma(y_2\theta x_2)]\,[\sigma(y_5\theta x_5)] = \prod_{i=1}^{5}[\sigma(y_i\theta x_i)]$$

$$\implies -\log(\Pr[\mathcal{D}|\{x_1, x_2, \ldots, x_5\}, \theta]) = \sum_{i=1}^{5} -\log(\sigma(y_i\theta x_i)) = \sum_{i=1}^{5} \log\left(1 + \exp(-y_i\theta x_i)\right)$$

The NLL defined above is referred to as the **logistic loss** and this model is referred to as (1-dimensional) **logistic regression**. Since we are classifying the coins as those that came up heads or tails, we are doing **binary classification**.

Logistic regression for binary classification is one of the most things in machine learning. E.g. Classifying whether a patient with feature x has cancer or not is another example where this procedure can be used.

Estimating the bias of multiple coins

In order to compute the MLE $\hat{\theta}$, we need to minimize the NLL on the previous slide, meaning that

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{5} \log\left(1 + \exp(-y_i \theta x_i)\right).$$

Unfortunately, this optimization problem can not be solved directly by taking derivatives like before. We need techniques from **numerical optimization** that studies efficiently minimizing complicated functions and the related computational properties.

If you like numerical optimization and want to see how to use it for solving difficult machine learning problems, you can take the CMPT 409 (Optimization for Machine Learning) course I am teaching in the Fall!

Once we have computed $\hat{\theta}$, we can use it to predict the probability of heads for a new coin that has feature x as $p = \sigma(\theta x)$.

Questions?

Generalizing to multiple dimensions

Suppose each coin i has a vector of features – its weight, air resistance, etc that affect its bias. If there are d such features, $x_i \in R^d$. Correspondingly, we will also have a vector of parameters $\theta \in \mathbb{R}^d$ and the linear model can be generalized as,

$$p_i = \sigma\left(\sum_{j=1}^{d} \theta_j x_{i,j}\right)$$

Writing this in terms of the dot product $\langle \theta, x_i \rangle = \sum_{j=1}^{d} \theta_j x_{i,j}$,

$$p_i = \frac{1}{1 + \exp(-\langle \theta, x_i \rangle)}$$

Suppose we toss the coins, and get the same $\mathcal{D} = HTHHT$, by following the same steps as before, the MLE can be given by:

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{5} \log\left(1 + \exp(-y_i \langle \theta x_i \rangle)\right).$$

Generalizing beyond coins

Suppose we are given n inputs in the form of their features (X) and labels y. Here, X is an $d \times n$-dimensional matrix such that the feature of input i is column $X_i \in \mathbb{R}^d$ and y is a n-dimensional vector such that $y_i \in \{-1, 1\}$.

In our coin example, the inputs were coins with different properties (features) and the labels corresponded to whether we got a heads ($y = 1$) or tails ($y = -1$). The feature-label could be characteristics of a patient and whether or not they have cancer, pixels in pictures of cats and dogs and whether it is a cat or a dog, text in the email and whether or not it is spam (Gmail uses a logistic regression model for classifying spam).

Using the linear model, $\Pr[y_i = 1] = p_i = \frac{1}{1+\exp(-\langle\theta, X_i\rangle)}$, we can write the general logistic regression **loss function** for binary classification and the corresponding MLE as its minimizer, i.e.

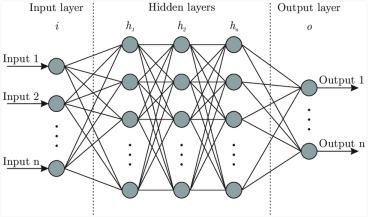
$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} \log\left(1 + \exp(-y_i \langle \theta X_i \rangle)\right).$$

This is the definition you will find in machine learning textbooks!

Generalizing beyond linear models

We are free to choose how to define p_i. We have been using a linear model such that, $p_i = \sigma(\langle \theta, X_i \rangle)$, but we could use *any* function $f(\theta, X_i)$ as an argument to the sigmoid function.

Designing such f functions is a major research direction. Current most popular models (used to classify videos on YouTube, rank posts on Facebook/Instagram) are (much) larger variants of **neural networks** that look like this:



If you found this lecture this fascinating, you can take the CMPT 410 (Machine Learning) course offered in the Fall!

Questions?