Distance Between Distributions

Here is an important question! How can you quantify how different two distributions are? In other words, if I have two Probability Mass Functions, can I calculate a number that says how much they diverge from one another? Here are three reasonable ways to quantify distance between distributions:

Total Variational Distance

Loop over all possible values and calculate the absolute difference in probability.

Definition: Total Variation Distance

Let X and Y be discrete random variables

$$\mathrm{TV}(X, Y) = \frac{1}{2} \sum_i |\,\mathrm{P}(X = i) - \mathrm{P}(Y = i)|$$

Earth Mover's Distance

Imagine one distribution is a lump of dirt. How much work would it take to make it look just like the other? Also called the Wasserstein metric.

This value is very meaningful, but it doesn't have a closed form equation. Instead it is computed by solving a linear program. If the range of values that the random variables can take on is of size n, the linear program has a run time of $O(n^3 \cdot \log n)$ which is very slow.

Kullback Leiber Divergence

Let X and Y be discrete random variables. Calculate the expected "excess surprise" from using Y as your Probability Mass Function instead of X when the actual Probability Mass Function is X.

Definition: Kullback Leiber Divergence

Let X and Y be discrete random variables

$$\begin{aligned}
\mathrm{KL}(X, Y) &= \sum_{x \in X} \mathrm{ExcessSurprise}(x) \cdot P(X = x) \\
&= \sum_{x \in X} \Big[\mathrm{Surprise}_X(x) - \mathrm{Surprise}_Y(x)\Big] \cdot P(X = x) \\
&= \sum_{x \in X} \Big[\log_2 \frac{1}{P(Y = x)} - \log_2 \frac{1}{P(X = x)}\Big] \cdot P(X = x) \\
&= \sum_{x \in X} -\log_2 P(Y = x) + \log_2 P(X = x) \cdot P(X = x) \\
&= \sum_{x \in X} \log_2 \frac{P(X = x)}{P(Y = x)} \cdot P(X = x)
\end{aligned}$$

Here is example code for calculating how different an observed pmf of hurricanes per year is from a predicted poisson distribution

```
from scipy import stats
import math

def kl_divergence(predicted_lambda, observed_pmf):
    """
    We predicted that the number of hurricanes would be
    X ~ Poisson(predicted_lambda) and observed a real world
    number of hurricanes Y ~ observed_pmf
    """
    X = stats.poisson(predicted_lambda)
    divergence = 0
    # loop over all the values of hurricanes
    for i in range(0, 40):
      pr_X_i = X.pmf(i)
      pr_Y_i = observed_pmf[i]
      excess_surprise_i = math.log(pr_X_i / pr_Y_i)
      divergence += excess_surprise_i * pr_X_i
    return divergence
```