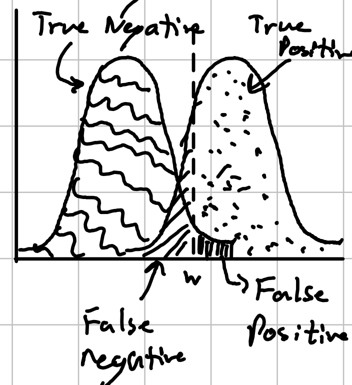
Lecture 5: TPR, FPR, Statistical Distributions, Central limit theorem, ROC/AUC, Model Comparisons

Dr. Farbin | 9/4/24

True vs. False in Statistics: Consider two distributions that overlap with one another. These distributions are based on histograms of data, close to the probability distribution. We choose as our boundary some w



- True Positive (TP): The number of positive instances correctly labled.
- False Positive (FP): The number of negative instances incorrectly labled.
- True Negative (TN): The number of negative instances correctly labled.
- False Negative (FN): The number of negative instances incorrectly labled.
- True Positive Rate (TPR):
  $TPR = TP/(TP+FN)$
- False Positive Rate (FPR):
  $FPR = FP/(FP+TN)$
- Accuracy (ACC):
  $ACC = (TP+TN)/(TP+FP+TN+FN)$

As your data's dimension increases, this gets harder.

Expectation Value $\mathbb{E}$: Given some random variable x and some probability distribution $P(x)$ such that $x \sim P(x)$, the expectation value of some function of x is

$$\mathbb{E}[f(x)] = \begin{cases} \int_{-\infty}^{\infty} f(x)\, p(x)\, dx & ; x \text{ is continnous} \\ \sum_{\{x_i\}} f(x_i) \cdot P(x_i) & ; x \text{ is discrete} \end{cases}$$

Mean, Variance, and Standard deviation: In the case where $f(x) = x$, we find that the mean of x is

$$\mu = \mathbb{E}[x] = \begin{cases} \int_{-\infty}^{\infty} x\, p(x)\, dx & ; x \text{ is continnous} \\ \sum_{\{x_i\}} x_i \cdot P(x_i) & ; x \text{ is discrete} \end{cases}$$

↳ Usually we do not have the prob. distribution, but we have the histogram(s) of the data. If $n(x)$ represents the count numbers for some given x, N represents the total number of data, and Δx is the bin size,

$$P(x) = \lim_{\substack{N \to \infty \\ \Delta x \to 0}} \frac{n(x)}{N \cdot \Delta x} \approx \frac{n(x)}{N \cdot \Delta x}$$

Thus we find

$$\mu = \mathbb{E}[x] \approx \frac{1}{N} \sum_{\substack{\text{all data} \\ \text{points}}} x_i$$

This gives us an estimate of the mean of x from the data itself

Variance measures the spread of your data and is given by

$$\sigma^2 = \mathbb{E}[x^2] - \mu^2 = \left[\frac{1}{N-1} \sum_{\{x_i\}} (x_i - \mu)^2\right]^2$$

While the standard deviation is given by

$$\sigma = \sqrt{\frac{1}{N-1} \sum_{\{x_i\}} (x_i - \mu)^2}$$

<u>Binomial distribution:</u> Consider the case of flipping a coin. We have N independent flips (a.k.a. tests) w/ boolean results (true for heads, false for tails). If you flipped 100 times, getting 52 H (true) and 48 T (false), how can you tell if the result makes sense or is due to an unfair coin?

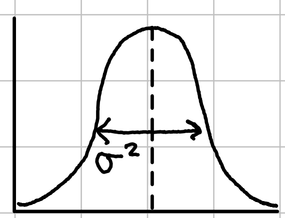
↳ Answering this simply requires one to consider

$$P(n \mid N, p) = [N! / (n!(N-n)!)] \cdot p^n (1-p)^{N-n}$$

where

n: # of successes (true values)

p: underlying prob. for success



For this type of distribution

$$\mathbb{E}(n) = N \cdot p$$

$$\sigma^2(p) = V(p) = N \cdot p(1-p)$$

Estimating using our data, we find

$$p \approx \frac{n}{N}$$

$$V(p) = \frac{p(1-p)}{N}$$

$$\mathrm{StD}(p) \propto \frac{1}{\sqrt{N}}$$

↳ This shows that the more trials and samples one collects, the better

Poisson distribution: Consider now the case where you own a store. You know the average # of customers. How many supplies and staff do you need to make sure you are prepared 90 % of the days?

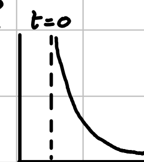
↳ We start with the Binomial distribution, but demand that $N \to \infty$ and $p \to 0$, so

$$\mathbb{E} \to \nu$$

$$f(n;\nu) = \frac{\nu^n}{n!} e^{-\nu}$$

$$\mathrm{StD}(n) = \sqrt{\nu}$$

Exponential Distribution: let's say you have a customer, how long do you need to wait before another arrives?



$$p(t|\tau) = \begin{cases} \tau e^{-t/\tau} & ; \ t \geq 0 \\ 0 & ; \ t < 0 \end{cases}$$

Distribution Summary: All in all, when considering a problem, one should find the underlying distribution, as the statistics associated with it can help mold one's approach in the analysis.

Central Limit Theorem: Consider two random variables x and y (not necessarily following the same distributions). The combined distribution will end us as a normal distribution

$$z = x + y$$

$$f(z;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2}$$

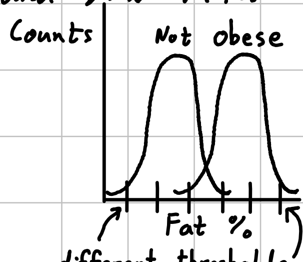
$$\mathbb{E}[z] = \mu$$

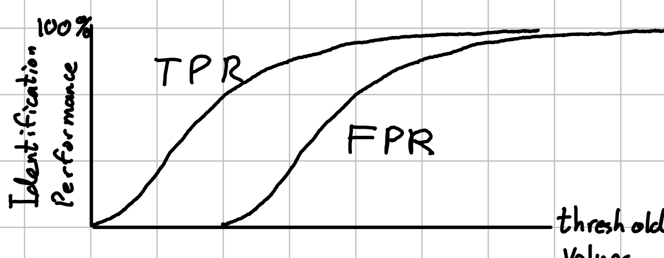
$$\mathrm{StD}[z] = \sigma$$

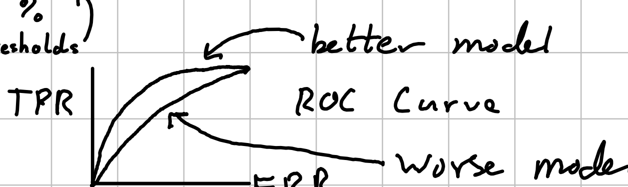
<u>Receiver Operating Characteristive Curve:</u> Consider a model identifying people as obese (positive) or not (negative). This model will assume that people below some fat % threshold value are not obese and vice versa. But how can we check the model's performance with different threshold choices?

↳ Answer: The ROC Curve, which graphically illustrates the performance of a classification model across different threshold settings.

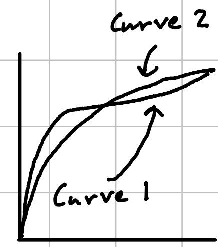
The ROC curve plots TPR (y-axis) against FPR (x-axis). A curve closer to the top left corner indicates a better performing model because it has high TPR and low FPR.







Area Under the Curve: Consider two ROC curves as depicted. Which one represents a better model? The Area under each curve, AUC, summarizes the overall performance of the classification model across all thresholds



↳ An AUC of 0.5 indicates a model with no discriminative ability, while an AUC of 1.0 represents a perfect model that correctly classifies all positive and negative instances.

Comparing Models: let's say you have two models based on the binomial distribution with 100 data points. We note

$$AUC(\text{Model } 1) = 0.75$$

$$AUC(\text{Model } 2) = 0.78$$

Is model 2 actually better?

↳ Since we are considering a binomial distribution, in order for model 2 to be statistically better than model 1, it needs to perform at least $\sqrt{100} \to 10\%$ better.

- In this case, the answer is no, but one should train and test both models multiple times, only then could one verify with confidence model 1 vs. 2