Exercise Walkthrough: Fish Classification

Justin Lanfermann

25. June 2025

Abstract

This document provides a detailed, step-by-step walkthrough of an exercise combining concepts from probability theory and basic machine learning. We will analyze a classification problem, derive formulas for data simulation using inverse transform sampling, calculate theoretical error rates for different classifiers, and discuss how to verify these results experimentally. Each step is explained with reference to the concepts from the "Discrete Probability Theory" script.

1 Problem Overview

We are tasked with classifying fish as either trout ($Y = 0$) or salmon ($Y = 1$) based on their measured size, a feature $X \in [0, 1]$. The problem provides us with the following model:

- **Prior Probabilities**: There are equal numbers of each fish type.

$$P(Y = 0) = P(Y = 1) = \frac{1}{2}$$

- **Likelihoods (Conditional PDFs)**: We know the distribution of sizes for each type of fish.

$$p(x|Y = 0) = 2 - 2x \quad \text{(Trout – smaller on average)}$$
$$p(x|Y = 1) = 2x \quad \text{(Salmon – larger on average)}$$

  Both densities are for $x \in [0, 1]$ and are zero otherwise.

Our goal is to use this information to build and evaluate simple classifiers.

2 Step 1: Inverse Transform Sampling Formulas

The first task is to find a way to generate synthetic data that follows our model. Specifically, we need to be able to sample a size X given a fish type Y. The standard method for this is **Inverse Transform Sampling** [1]. This technique allows us to generate random samples from a distribution if we know its Cumulative Distribution Function (CDF) [2].

The core idea is: if U is a random variable uniformly distributed on $[0, 1]$, then $X = F_X^{-1}(U)$ is a random variable with CDF F_X. We will apply this to our conditional distributions.

2.1 Sampling a Trout's Size (Y=0)

First, we need the conditional CDF for a trout's size, $F_{X|Y=0}(x)$, by integrating the PDF $p(x|Y=0)$.

$$\begin{aligned} F_{X|Y=0}(x) = P(X \le x|Y=0) &= \int_0^x p(z|Y=0)\,dz \\ &= \int_0^x (2-2z)\,dz \\ &= \left[2z - z^2\right]_0^x \\ &= 2x - x^2 \quad \text{for } x \in [0,1] \end{aligned}$$

Now, we set this equal to a uniform random variable $u \in [0,1]$ and solve for x:

$$\begin{aligned} 2x - x^2 &= u \\ x^2 - 2x + u &= 0 \\ x^2 - 2x + 1 &= 1 - u \qquad \text{(Completing the square)} \\ (x-1)^2 &= 1 - u \\ x - 1 &= \pm\sqrt{1-u} \\ x &= 1 \pm \sqrt{1-u} \end{aligned}$$

Since we know x must be in the interval $[0,1]$, we must choose the negative sign. If we chose the positive sign, x would be ≥ 1. Thus, the inverse transformation is:

$$x = F^{-1}_{X|Y=0}(u) = 1 - \sqrt{1-u}$$

2.2 Sampling a Salmon's Size (Y=1)

We follow the same procedure for the salmon's size distribution.

$$\begin{aligned} F_{X|Y=1}(x) = P(X \le x|Y=1) &= \int_0^x p(z|Y=1)\,dz \\ &= \int_0^x 2z\,dz \\ &= \left[z^2\right]_0^x \\ &= x^2 \quad \text{for } x \in [0,1] \end{aligned}$$

Again, we set this equal to u and solve for x:

$$\begin{aligned} x^2 &= u \\ x &= \sqrt{u} \end{aligned}$$

We take the positive root since $x \in [0,1]$. So, the inverse transformation is:

$$x = F^{-1}_{X|Y=1}(u) = \sqrt{u}$$

Tutor's Note. *These two formulas are the key to the simulation part of the exercise. To generate a data point (X,Y), you first flip a coin to decide if it's a trout or salmon (to determine Y), and then use the corresponding formula with a new random number u to generate its size X.*

3 Step 2: Implementation and Verification

This part of the exercise asks for a Python implementation. Here, we'll outline the logic. A function `generate_data(N)` would perform the following steps `N` times:

1. **Generate Y**: Draw a random number $v \sim \text{Unif}(0, 1)$. If $v < 0.5$, set $Y = 0$ (trout). Otherwise, set $Y = 1$ (salmon). This respects the prior probabilities $P(Y = 0) = P(Y = 1) = 1/2$.
2. **Generate X**: Draw another random number $u \sim \text{Unif}(0, 1)$.
   - If $Y = 0$, calculate $X = 1 - \sqrt{1 - u}$.
   - If $Y = 1$, calculate $X = \sqrt{u}$.
3. Store the pair (X, Y).

After generating a large number of samples (e.g., $N = 10000$), you can verify your formulas. You would split the data by class (all (X, Y) pairs with $Y = 0$ and all with $Y = 1$). Then, for each class, you can plot the Empirical CDF (ECDF) of the generated X values and overlay it with the theoretical CDF formula we derived in Section 2. The ECDF and the theoretical curve should match very closely, confirming our derivations are correct.

4 Step 3: Theoretical Error Rate Analysis

Now we calculate the average error rate for each classifier "on paper". The average error rate is the probability that the classifier's prediction $\hat{y}$ is not equal to the true class Y, i.e., $P(\hat{y} \neq Y)$. This can be calculated as the expectation of an indicator function [3], $E[\chi_{\{\hat{y} \neq Y\}}]$. We use the **Law of Total Probability** [4] to break this down by conditioning on the true class Y:

$$\begin{aligned}\text{Error Rate} &= P(\hat{y} \neq Y) \\ &= P(\hat{y} \neq Y|Y = 0)P(Y = 0) + P(\hat{y} \neq Y|Y = 1)P(Y = 1) \\ &= \frac{1}{2}\left(P(\hat{y} \neq 0|Y = 0) + P(\hat{y} \neq 1|Y = 1)\right)\end{aligned}$$

4.1 Classifier A: $\hat{y} = 0$ if $x \leq x_t$, $\hat{y} = 1$ if $x > x_t$

- **Error on Trout (Y=0)**: We predict 1 ($\hat{y} = 1$) when the truth is 0. This happens when $X > x_t$. The probability is $P(X > x_t|Y = 0)$.
- **Error on Salmon (Y=1)**: We predict 0 ($\hat{y} = 0$) when the truth is 1. This happens when $X \leq x_t$. The probability is $P(X \leq x_t|Y = 1)$.

Let's calculate these probabilities by integrating the PDFs over the error regions:

$$P(X > x_t|Y = 0) = \int_{x_t}^{1} (2 - 2x)\, dx = [2x - x^2]_{x_t}^{1} = (2 - 1) - (2x_t - x_t^2) = (1 - x_t)^2$$

$$P(X \leq x_t|Y = 1) = \int_{0}^{x_t} 2x\, dx = [x^2]_0^{x_t} = x_t^2$$

The total error rate is:

$$\begin{aligned}\text{Error}_A(x_t) &= \frac{1}{2}\left((1 - x_t)^2 + x_t^2\right) \\ &= \frac{1}{2}(1 - 2x_t + x_t^2 + x_t^2) \\ &= x_t^2 - x_t + \frac{1}{2}\end{aligned}$$

To find the optimal threshold x_t that minimizes this error, we take the derivative with respect to x_t and set it to zero:

$$\frac{d}{dx_t}\text{Error}_A(x_t) = 2x_t - 1 = 0 \implies x_t = \frac{1}{2}$$

The minimum possible average error rate for this classifier is at $x_t = 1/2$, which is:

$$\text{Error}_A(1/2) = \left(\frac{1}{2}\right)^2 - \frac{1}{2} + \frac{1}{2} = \frac{1}{4}$$

4.2 Classifier B: "Reverse of A"

Here, $\hat{y} = 1$ if $x \leq x_t$ and $\hat{y} = 0$ if $x > x_t$.

- **Error on Trout (Y=0)**: We predict 1 when $X \leq x_t$. Prob: $P(X \leq x_t | Y = 0)$.
- **Error on Salmon (Y=1)**: We predict 0 when $X > x_t$. Prob: $P(X > x_t | Y = 1)$.

$$P(X \leq x_t | Y = 0) = F_{X|Y=0}(x_t) = 2x_t - x_t^2$$
$$P(X > x_t | Y = 1) = 1 - F_{X|Y=1}(x_t) = 1 - x_t^2$$

The total error rate is:

$$\text{Error}_B(x_t) = \frac{1}{2}(2x_t - x_t^2 + 1 - x_t^2) = -x_t^2 + x_t + \frac{1}{2}$$

This classifier is clearly worse. Its error is minimized at the boundaries ($x_t = 0$ or $x_t = 1$) with an error of $1/2$, and maximized at $x_t = 1/2$ with an error of $3/4$.

4.3 Classifier C: "Guessing"

Here, we flip a fair coin for the prediction $\hat{y} \sim \text{Ber}(1/2)$, independent of X and Y.

$$\begin{aligned}\text{Error}_C &= P(\hat{y} \neq Y) = P(\hat{y} = 1, Y = 0) + P(\hat{y} = 0, Y = 1) \\ &= P(\hat{y} = 1)P(Y = 0) + P(\hat{y} = 0)P(Y = 1) \qquad \text{(by independence)} \\ &= \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) + \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}\end{aligned}$$

The error rate is a constant $1/2$, regardless of the threshold x_t.

4.4 Classifier D: "Everything's a salmon"

We always predict $\hat{y} = 1$. An error occurs if and only if the fish is actually a trout ($Y = 0$).

$$\text{Error}_D = P(\hat{y} \neq Y) = P(1 \neq Y) = P(Y = 0) = \frac{1}{2}$$

The error rate is also a constant $1/2$.

5 Step 4: Experimental Confirmation

This final step again involves coding. The goal is to see if our theoretical error rates from Section 4 hold up in practice.

Procedure:

1. For each specified threshold $x_t \in \{0.05, 0.1, \ldots, 0.95\}$:
2. Run a block of 10 experiments. In each experiment:
   (a) Generate a new, large dataset of size N (e.g., $N = 1000$ or $N = 10000$) using the function from Section 3.
   (b) For each classifier (A, B, C, D), apply its rule to the generated sizes X to get predictions $\hat{Y}$.
   (c) Calculate the empirical error rate for this single dataset: $\frac{1}{N}\sum_{i=1}^{N} \chi_{\{\hat{y}_i \neq y_i\}}$.
3. After the 10 experiments, you have 10 empirical error rates for this x_t. Calculate their mean and standard deviation.
4. Repeat for both $N = 1000$ and $N = 10000$.

Expected Results and Analysis: You should produce plots showing the theoretical error curves (e.g., the parabola $x_t^2 - x_t + 1/2$ for Classifier A) and overlay the mean empirical error rates for each x_t. The standard deviations can be shown as error bars.

A key observation relates to the standard deviation of the estimates. The empirical error rate is an average of N i.i.d. Bernoulli trials (where success is a classification error). According to the **Central Limit Theorem (CLT)** [5], the standard deviation of such an average (known as the standard error) is proportional to $1/\sqrt{N}$.

Therefore, when you increase the sample size from $N = 1000$ to $N = 10000$ (a factor of 10), you should expect the standard deviation of your error rate estimates to decrease by a factor of $\sqrt{10} \approx 3.16$. This demonstrates a fundamental principle of statistics: more data leads to more precise estimates.

In-depth Concepts

[1] Inverse Transform Sampling (based on Prop. 2.23, Rem. 2.24): This is a powerful method to generate random numbers from any probability distribution, given its Cumulative Distribution Function (CDF), $F_X(x)$. The theorem states that if U is a random variable that is uniformly distributed on the interval $[0, 1]$, then the random variable $X = F_X^{-1}(U)$ has the CDF F_X. This works because $P(X \leq x) = P(F_X^{-1}(U) \leq x) = P(U \leq F_X(x))$. Since U is uniform on $[0, 1]$, the probability that it's less than or equal to some value $v \in [0, 1]$ is simply v. Thus, $P(U \leq F_X(x)) = F_X(x)$, which proves that the generated X has the correct CDF.

[2] Cumulative Distribution Function (CDF) (Def. 1.21): The CDF of a random variable X, denoted $F_X(x)$, gives the probability that X will take a value less than or equal to x. Formally, $F_X(x) = P(X \leq x)$. For a continuous random variable with a Probability Density Function (PDF) $p_X(z)$, the CDF is found by integration: $F_X(x) = \int_{-\infty}^{x} p_X(z)\, dz$. The CDF "accumulates" probability, starting at 0 for $-\infty$ and ending at 1 for $+\infty$. It provides a complete description of the distribution.

[3] Expectation of a Function of Random Variables (LOTUS, Lem. 2.2): The "Law of the Unconscious Statistician" (LOTUS) provides a direct way to calculate the expected value of a function of a random variable, $g(X)$, without first finding the distribution of $Y = g(X)$. For a continuous random variable X with PDF $p_X(x)$, the formula is $E[g(X)] = \int_{-\infty}^{\infty} g(x)p_X(x)\, dx$. In our exercise, the "function" is the indicator of a classification error, $\chi_{\{\hat{y}(X) \neq Y\}}$, and we compute its expectation to find the error probability.

[4] Law of Total Probability (Thm. 1.65): This law allows us to find the probability of an event A by considering a set of mutually exclusive and exhaustive events (a partition) $\{B_1, B_2, \dots\}$. The formula is $P(A) = \sum_i P(A|B_i)P(B_i)$. It's a "divide and conquer" strategy for probabilities. In our error analysis, we partition the space by the true class of the fish, $\{Y = 0, Y = 1\}$, to compute the overall error rate. The law of total expectation (or tower property, Thm. 2.32) is the analogous rule for expectations: $E[X] = E[E[X|Y]]$.

[5] Central Limit Theorem (CLT) (Thm. 2.64): The CLT is one of the most important results in probability. It states that, under general conditions, the sum (or average) of a large number of independent and identically distributed (i.i.d.) random variables will be approximately normally distributed, regardless of the original distribution of the variables. The empirical error rate in our experiment is the average of N i.i.d. Bernoulli variables (1 for an error, 0 for correct). The CLT explains why the distribution of this average becomes more concentrated (has a smaller standard deviation) as N increases, and specifically why the standard deviation scales with $1/\sqrt{N}$.