Mathematical Foundation of Machine Learning Assignment V

Shedrack Eze
Andrew ID: seze

Question 1: Gradient Descent Analysis

Consider optimizing the function $f(x) = x^4 + 2x^2$ using gradient descent.

(a) Calculating the first derivative $f'(x)$ and second derivative $f''(x)$ of the function. Determine whether $f(x)$ is convex for all $x \in \mathbb{R}$ by analyzing its second derivative.

SOLUTION

Step 1: Computing first derivative using power rule

$$f(x) = x^4 + 2x^2$$

$$\frac{d}{dx}(x^n) = n \cdot x^{n-1}$$

For x^4:

$$\frac{d}{dx}(x^4) = 4 \cdot x^{4-1} = 4x^3$$

For $2x^2$:

$$\frac{d}{dx}(2x^2) = 2 \cdot \frac{d}{dx}(x^2) = 2 \cdot (2 \cdot x^{2-1}) = 2 \cdot (2x) = 4x$$

Combine terms:

$$f'(x) = 4x^3 + 4x$$

Factor:

$$f'(x) = 4x(x^2 + 1)$$

Step 2: Calculating the second derivative

$$f'(x) = 4x^3 + 4x$$

For $4x^3$:

$$\frac{d}{dx}(4x^3) = 4 \cdot \frac{d}{dx}(x^3) = 4 \cdot (3 \cdot x^{3-1}) = 4 \cdot (3x^2) = 12x^2$$

For $4x$:

$$\frac{d}{dx}(4x) = 4 \cdot \frac{d}{dx}(x) = 4 \cdot (1 \cdot x^{1-1}) = 4 \cdot (1 \cdot x^0) = 4 \cdot 1 = 4$$

Combine terms:

$$f''(x) = 12x^2 + 4$$

Step 3: Convexity analysis
A twice-differentiable function is convex if $f''(x) \geq 0$ for all $x \in \mathbb{R}$.

Test for $x > 0$:

Let x be any positive real number ($x > 0$):

$$x^2 > 0 \quad \text{(positive} \times \text{positive} = \text{positive)}$$

$$12x^2 > 0 \quad \text{(positive} \times \text{positive} = \text{positive)}$$

$$12x^2 + 4 > 0 + 4 = 4 > 0$$

Therefore: $f''(x) > 0$ for all $x > 0$

Test for $x < 0$:

Let x be any negative real number ($x < 0$):

$$x^2 > 0 \quad \text{(negative} \times \text{negative} = \text{positive)}$$

$$12x^2 > 0 \quad \text{(positive} \times \text{positive} = \text{positive)}$$

$$12x^2 + 4 > 0 + 4 = 4 > 0$$

Therefore: $f''(x) > 0$ for all $x < 0$

Test for $x = 0$:

$$x^2 = 0$$

$$12x^2 = 12 \times 0 = 0$$

$$12x^2 + 4 = 0 + 4 = 4 > 0$$

Therefore: $f''(0) > 0$

Conclusion: Since $f''(x) > 0$ for all $x \in \mathbb{R}$, and in particular $f''(x) \geq 0$ for all $x \in \mathbb{R}$, the function $f(x)$ is strictly convex for all $x \in \mathbb{R}$.

$$\boxed{f'(x) = 4x^3 + 4x, \quad f''(x) = 12x^2 + 4, \quad \text{Convex: It is Convex}}$$

(b) Perform three iterations of gradient descent starting from $x_0 = 1.5$ with learning rate $\eta = 0.1$.

Step 1: Recall the gradient descent update rule and gradient form

$$x_{i+1} = x_i - \eta \cdot f'(x_i)$$

$$f'(x) = 4x(x^2 + 1)$$

Iteration 0: $i = 0$, $x_0 = 1.5$

(i) Current position:

$$x_0 = 1.5$$

(ii) Gradient calculation:

$$f'(1.5) = 4x(x^2 + 1) = 4 \times 1.5 \times ((1.5)^2 + 1)$$

$$(1.5)^2 = 1.5 \times 1.5 = 2.25$$

$$(1.5)^2 + 1 = 2.25 + 1 = 3.25$$

$$4 \times 1.5 = 6$$

$$6 \times 3.25 = 19.5$$

$$f'(1.5) = 19.5$$

(iii) Update calculation:

$$x_1 = x_0 - \eta \cdot f'(x_0) = 1.5 - 0.1 \times 19.5$$

$$0.1 \times 19.5 = 1.95$$

$$x_1 = 1.5 - 1.95 = -0.45$$

(iv) Function value:

$$f(x) = x^4 + 2x^2$$

$$f(1.5) = (1.5)^4 + 2(1.5)^2$$

$$(1.5)^2 = 1.5 \times 1.5 = 2.25$$

$$(1.5)^4 = (1.5)^2 \times (1.5)^2 = 2.25 \times 2.25 = 5.0625$$

$$2(1.5)^2 = 2 \times 2.25 = 4.5$$

$$f(1.5) = 5.0625 + 4.5 = 9.5625$$

Iteration 1: $i = 1$, $x_1 = -0.45$

(i) Current position:

$$x_1 = -0.45$$

(ii) Gradient calculation:

$$f'(-0.45) = 4x(x^2 + 1) = 4 \times (-0.45) \times ((-0.45)^2 + 1)$$

$$(-0.45)^2 = (-0.45) \times (-0.45) = 0.2025$$

$$(-0.45)^2 + 1 = 0.2025 + 1 = 1.2025$$

$$4 \times (-0.45) = -1.8$$

$$-1.8 \times 1.2025 = -2.1645$$

$$f'(-0.45) = -2.1645$$

(iii) Update calculation:

$$x_2 = x_1 - \eta \cdot f'(x_1) = -0.45 - 0.1 \times (-2.1645)$$

$$0.1 \times (-2.1645) = -0.21645$$

$$x_2 = -0.45 - (-0.21645) = -0.45 + 0.21645 = -0.23355$$

(iv) Function value:

$$f(x) = x^4 + 2x^2$$

$$f(-0.45) = (-0.45)^4 + 2(-0.45)^2$$

$$(-0.45)^2 = (-0.45) \times (-0.45) = 0.2025$$

$$(-0.45)^4 = (-0.45)^2 \times (-0.45)^2 = 0.2025 \times 0.2025 = 0.04100625$$

$$2(-0.45)^2 = 2 \times 0.2025 = 0.405$$

$$f(-0.45) = 0.04100625 + 0.405 = 0.44600625$$

Iteration 2: $i = 2$**,** $x_2 = -0.23355$

(i) Current position:

$$x_2 = -0.23355$$

(ii) Gradient calculation:

$$f'(-0.23355) = 4x(x^2 + 1) = 4 \times (-0.23355) \times ((-0.23355)^2 + 1)$$

$$(-0.23355)^2 = (-0.23355) \times (-0.23355) = 0.054545$$

$$(-0.23355)^2 + 1 = 0.054545 + 1 = 1.054545$$

$$4 \times (-0.23355) = -0.9342$$

$$-0.9342 \times 1.054545 = -0.985176$$

$$f'(-0.23355) = -0.985176$$

(iii) Update calculation:

$$x_3 = x_2 - \eta \cdot f'(x_2) = -0.23355 - 0.1 \times (-0.985176)$$

$$0.1 \times (-0.985176) = -0.0985176$$

$$x_3 = -0.23355 - (-0.0985176) = -0.23355 + 0.0985176 = -0.1350324$$

(iv) Function value:

$$f(x) = x^4 + 2x^2$$

$$f(-0.23355) = (-0.23355)^4 + 2(-0.23355)^2$$

$$(-0.23355)^2 = (-0.23355) \times (-0.23355) = 0.054545$$

$$(-0.23355)^4 = (-0.23355)^2 \times (-0.23355)^2 = 0.054545 \times 0.054545 = 0.002975$$

$$2(-0.23355)^2 = 2 \times 0.054545 = 0.10909$$

$$f(-0.23355) = 0.002975 + 0.10909 = 0.112065$$

(c) Now perform one iteration from the same starting point $x_0 = 1.5$ but with learning rate $\eta = 0.5$.

Step 1: Use the same starting point

$$x_0 = 1.5$$

Step 2: Gradient calculation

$$f'(1.5) = 4x(x^2 + 1) = 4 \times 1.5 \times ((1.5)^2 + 1)$$

$$(1.5)^2 = 1.5 \times 1.5 = 2.25$$
$$(1.5)^2 + 1 = 2.25 + 1 = 3.25$$
$$4 \times 1.5 = 6$$
$$6 \times 3.25 = 19.5$$
$$f'(1.5) = 19.5$$

Step 3: Update with larger learning rate

$$x_1 = x_0 - \eta \cdot f'(x_0) = 1.5 - 0.5 \times 19.5$$

$$0.5 \times 19.5 = 9.75$$
$$x_1 = 1.5 - 9.75 = -8.25$$

Step 4: Function value at new point

$$f(x) = x^4 + 2x^2$$

$$f(-8.25) = (-8.25)^4 + 2(-8.25)^2$$
$$(-8.25)^2 = (-8.25) \times (-8.25) = 68.0625$$
$$(-8.25)^4 = (-8.25)^2 \times (-8.25)^2 = 68.0625 \times 68.0625 = 4632.12890625$$
$$2(-8.25)^2 = 2 \times 68.0625 = 136.125$$
$$f(-8.25) = 4632.12890625 + 136.125 = 4768.25390625$$

Step 5: Comparison and analysis

- With $\eta = 0.1$: $x_1 = -0.45$, $f(x_1) = 0.446$ (decreased from 9.5625)
- With $\eta = 0.5$: $x_1 = -8.25$, $f(x_1) = 4768.254$ (increased dramatically)

Explanation: The large learning rate causes overshooting. The gradient at $x_0 = 1.5$ is large (19.5), and multiplying by 0.5 gives a step size of 9.75, which moves us past the minimum at $x = 0$ to a point where the function value is much higher.

$$\boxed{\text{Large } \eta \text{ causes overshoot and divergence in this convex function}}$$

(d) The condition number of the Hessian matrix affects gradient descent convergence. Calculate the Hessian (second derivative) of $f(x)$ at $x = 1$ and discuss how its value might influence the convergence behavior of gradient descent.

Step 1: Recall Hessian for 1D function

For a single-variable function, the Hessian is just the second derivative:

$$H(x) = f''(x) = 12x^2 + 4$$

Step 2: Calculate at $x = 1$

$$H(1) = 12(1)^2 + 4 = 12 \times 1 + 4 = 12 + 4 = 16$$

Step 3: Discuss condition number influence

The condition number κ is defined as:

$$\kappa = \frac{\max_{x \in \text{domain}} f''(x)}{\min_{x \in \text{domain}} f''(x)}$$

For our function $f''(x) = 12x^2 + 4$:

- **Minimum value:** When $x = 0$, $f''(0) = 4$
- **Maximum value:** As $|x| \to \infty$, $f''(x) \to \infty$

Therefore, over the entire real line:

$$\kappa = \frac{\infty}{4} = \infty$$

Even for a bounded domain, say $x \in [-a, a]$:

$$\kappa = \frac{12a^2 + 4}{4} = 3a^2 + 1$$

Step 4: Influence on gradient descent

A large condition number indicates that the function's curvature varies significantly across different regions. This variation presents a challenge for gradient descent optimization because it necessitates the use of very small learning rates to prevent divergence in areas where the curvature is particularly high. The maximum allowable learning rate is constrained by the region of steepest curvature, following the rule $\eta < \frac{2}{L}$, where L represents the maximum curvature encountered. At the specific point $x = 1$, the curvature measures 16, which is moderately high and therefore requires careful selection of the learning rate to ensure stable convergence during optimization.

$$\boxed{f''(1) = 16, \quad \text{Large variation in } f''(x) \text{ over large domain slows GD}}$$

Question 2: Constrained Optimization & KKT Conditions

Consider the optimization problem:

$$\min_{x_1,x_2} f(x_1, x_2) = x_1^2 + x_2^2$$

subject to the constraint:

$$g(x_1, x_2) = x_1 + x_2 - 1 \leq 0$$

(a) Lagrangian Function and KKT Conditions

Step 1: Write the Lagrangian function

We have the optimization problem:

- Objective function: $f(x_1, x_2) = x_1^2 + x_2^2$
- Constraint: $g(x_1, x_2) = x_1 + x_2 - 1 \leq 0$

The Lagrangian function combines the objective function with the constraint using a Lagrange multiplier λ:

$$L(x_1, x_2, \lambda) = f(x_1, x_2) + \lambda \cdot g(x_1, x_2)$$

Substituting the specific functions:

$$L(x_1, x_2, \lambda) = x_1^2 + x_2^2 + \lambda(x_1 + x_2 - 1)$$

Step 2: Stating the KKT conditions

For a minimization problem with inequality constraint $g(x) \leq 0$, the Karush-Kuhn-Tucker conditions are:

[label=(0)]

1. **Primal feasibility:** The solution must satisfy the original constraint:

$$g(x_1, x_2) \leq 0$$

2. **Dual feasibility:** The Lagrange multiplier must be non-negative:

$$\lambda \geq 0$$

3. **Complementary slackness:** Either the constraint is active (equality holds) or the Lagrange multiplier is zero:

$$\lambda \cdot g(x_1, x_2) = 0$$

4. **Stationarity:** The gradient of the Lagrangian with respect to the primal variables must be zero:

$$\nabla_{x_1,x_2} L(x_1, x_2, \lambda) = 0$$

Which expands to:

$$\nabla f(x_1, x_2) + \lambda \nabla g(x_1, x_2) = 0$$

(b) Solving the KKT Conditions

Step 1: Apply the stationarity condition

The stationarity condition requires:

$$\frac{\partial L}{\partial x_1} = 0 \quad \text{and} \quad \frac{\partial L}{\partial x_2} = 0$$

Compute partial derivatives:

$$\frac{\partial L}{\partial x_1} = \frac{\partial}{\partial x_1}[x_1^2 + x_2^2 + \lambda(x_1 + x_2 - 1)] = 2x_1 + \lambda$$

$$\frac{\partial L}{\partial x_2} = \frac{\partial}{\partial x_2}[x_1^2 + x_2^2 + \lambda(x_1 + x_2 - 1)] = 2x_2 + \lambda$$

Set derivatives to zero:

$$2x_1 + \lambda = 0 \quad \text{(Equation 1)}$$

$$2x_2 + \lambda = 0 \quad \text{(Equation 2)}$$

Step 2: Solve the stationarity equations

From Equation 1:

$$2x_1 + \lambda = 0 \Rightarrow x_1 = -\frac{\lambda}{2}$$

From Equation 2:

$$2x_2 + \lambda = 0 \Rightarrow x_2 = -\frac{\lambda}{2}$$

Therefore:

$$x_1 = x_2$$

Step 3: Consider the complementary slackness condition

The complementary slackness condition is:

$$\lambda \cdot g(x_1, x_2) = 0$$

This gives us two cases to consider:

Case 1: $\lambda = 0$

If $\lambda = 0$, then from Equations 1 and 2:

$$2x_1 + 0 = 0 \Rightarrow x_1 = 0$$

$$2x_2 + 0 = 0 \Rightarrow x_2 = 0$$

Then we check primal feasibility:

$$g(0, 0) = 0 + 0 - 1 = -1 \leq 0$$

The constraint is satisfied.

So we have a solution: $(x_1, x_2, \lambda) = (0, 0, 0)$

Case 2: $\lambda > 0$ **and** $g(x_1, x_2) = 0$

If $\lambda > 0$, then complementary slackness requires:

$$g(x_1, x_2) = x_1 + x_2 - 1 = 0$$

But from Step 2, we know $x_1 = x_2$, so:

$$x_1 + x_2 - 1 = 0 \ \Rightarrow \ 2x_1 = 1 \ \Rightarrow \ x_1 = \frac{1}{2}, \ x_2 = \frac{1}{2}$$

find λ from Equation 1:

$$2\left(\frac{1}{2}\right) + \lambda = 0 \Rightarrow 1 + \lambda = 0 \Rightarrow \lambda = -1$$

But this violates dual feasibility ($\lambda \geq 0$).

Step 4: Check the boundary case
Let's check if $\lambda = 0$ with $g(x_1, x_2) = 0$ is possible:

If $g(x_1, x_2) = 0$ and $x_1 = x_2$, then:

$$x_1 + x_1 - 1 = 0 \Rightarrow x_1 = \frac{1}{2}, \quad x_2 = \frac{1}{2}$$

But from stationarity with $\lambda = 0$:

$$2\left(\frac{1}{2}\right) + 0 = 1 \neq 0$$

This also violates the stationarity condition.

Step 5: The final solution

The only solution that satisfies all KKT conditions is:

$$x_1^* = 0, \quad x_2^* = 0, \quad \lambda^* = 0$$

(c) Verification of KKT Conditions

Verification 1: Primal feasibility We need to verify that $g(x_1^*, x_2^*) \leq 0$:

$$g(0, 0) = 0 + 0 - 1 = -1$$

Since $-1 \leq 0$, the primal feasibility condition is satisfied.

Verification 2: Dual feasibility

We need to verify that $\lambda^* \geq 0$:

$\lambda^* = 0 \geq 0$, so the dual feasibility condition is satisfied.

Verification 3: Complementary slackness

We need to verify that $\lambda^* \cdot g(x_1^*, x_2^*) = 0$:

$$\lambda^* \cdot g(0,0) = 0 \cdot (-1) = 0$$

The complementary slackness condition is satisfied.

Verification 4: Stationarity condition

We need to verify that $\nabla f(x_1^*, x_2^*) + \lambda^* \nabla g(x_1^*, x_2^*) = 0$:

Compute $\nabla f(x_1, x_2)$:

$$\nabla f(x_1, x_2) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix}$$

At $(0,0)$:

$$\nabla f(0,0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Compute $\nabla g(x_1, x_2)$:

$$\nabla g(x_1, x_2) = \begin{bmatrix} \frac{\partial g}{\partial x_1} \\ \frac{\partial g}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Now compute:

$$\nabla f(0,0) + \lambda^* \nabla g(0,0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} + 0 \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

The stationarity condition is satisfied.

Geometric interpretation:

The objective function $f(x_1, x_2) = x_1^2 + x_2^2$ represents concentric circles centered at the origin. The minimum of this function is at $(0,0)$ with value 0. The constraint $x_1 + x_2 \leq 1$ defines a half-plane. The line $x_1 + x_2 = 1$ is the boundary of this feasible region. Since the unconstrained minimum $(0,0)$ lies inside the feasible region (as $0+0-1 = -1 \leq 0$), the constraint is not active at the optimum. This is why the Lagrange multiplier $\lambda^* = 0$ - the constraint doesn't affect the solution because we can reach the unconstrained minimum without violating the constraint.

Question 3: Probability Foundations

A machine learning system for fraud detection has the following characteristics:

- Probability that a transaction is fraudulent: $P(F) = 0.01$
- If a transaction is fraudulent, probability the system flags it: $P(A|F) = 0.95$
- If a transaction is legitimate, probability the system flags it: $P(A|\neg F) = 0.02$

(a) Calculate $P(A)$ using Law of Total Probability

Step 1: Identify given probabilities

$$\begin{aligned} P(F) &= 0.01 \quad \text{(probability transaction is fraudulent)} \\ P(A|F) &= 0.95 \quad \text{(probability system flags given fraudulent)} \\ P(A|\neg F) &= 0.02 \quad \text{(probability system flags given legitimate)} \end{aligned}$$

Step 2: Calculate complementary probabilities

$$P(\neg F) = 1 - P(F) = 1 - 0.01 = 0.99$$

Step 3: Apply Law of Total Probability

$$P(A) = P(A|F) \cdot P(F) + P(A|\neg F) \cdot P(\neg F)$$

$$P(A) = (0.95 \times 0.01) + (0.02 \times 0.99)$$

$$P(A) = 0.0095 + 0.0198 = 0.0293$$

$$\boxed{P(A) = 0.0293}$$

(b) Calculate conditional probabilities using Bayes' Theorem

(i) $P(F|A)$ - probability fraudulent given flagged

Step 1: Applying Bayes' Theorem

$$P(F|A) = \frac{P(A|F) \cdot P(F)}{P(A)}$$

$$P(F|A) = \frac{0.95 \times 0.01}{0.0293}$$

$$P(F|A) = \frac{0.0095}{0.0293} \approx 0.3242$$

$$\boxed{P(F|A) \approx 0.3242}$$

(ii) $P(\neg F|A)$ **- The probability of legitimate being flagged**
Step 1: Apply Bayes' Theorem

$$P(\neg F|A) = \frac{P(A|\neg F) \cdot P(\neg F)}{P(A)}$$

$$P(\neg F|A) = \frac{0.02 \times 0.99}{0.0293}$$

$$P(\neg F|A) = \frac{0.0198}{0.0293} \approx 0.6758$$

Step 2: Verify probabilities sum to 1

$$P(F|A) + P(\neg F|A) = 0.3242 + 0.6758 = 1$$

$$\boxed{P(\neg F|A) \approx 0.6758}$$

(c) System Usefulness Analysis

(i) Check if $P(F|A) > 0.5$

We calculated:

$$P(F|A) \approx 0.3242$$

Since $0.3242 < 0.5$, the system **does not meet** the usefulness criterion.
(ii) Find required $P(A|F)$ **for** $P(F|A) = 0.5$
Step 1: Set up equation using Bayes' Theorem

$$P(F|A) = \frac{P(A|F) \cdot P(F)}{P(A)} = 0.5$$

Step 2: Express $P(A)$ in terms of $P(A|F)$

$$P(A) = P(A|F) \cdot P(F) + P(A|\neg F) \cdot P(\neg F)$$

$$P(A) = P(A|F) \times 0.01 + 0.02 \times 0.99$$
$$P(A) = 0.01 \cdot P(A|F) + 0.0198$$

Step 3: Substitute into Bayes' Theorem

$$\frac{P(A|F) \times 0.01}{0.01 \cdot P(A|F) + 0.0198} = 0.5$$

Step 4: Solve for $P(A|F)$

$$P(A|F) \times 0.01 = 0.5 \times (0.01 \cdot P(A|F) + 0.0198)$$

$$0.01 \cdot P(A|F) = 0.005 \cdot P(A|F) + 0.0099$$
$$0.01 \cdot P(A|F) - 0.005 \cdot P(A|F) = 0.0099$$
$$0.005 \cdot P(A|F) = 0.0099$$
$$P(A|F) = \frac{0.0099}{0.005} = 1.98$$

This is impossible since probabilities cannot exceed 1.

Step 5: Interpret the result
Since we got $P(A|F) > 1$, it means that with $P(F) = 0.01$ and $P(A|\neg F) = 0.02$, it's **impossible** to achieve $P(F|A) = 0.5$.

(iii) Practical implications

The analysis shows some practical challenges for deploying this fraud detection system. A major issue is the **base rate fallacy** because despite the system's high accuracy (95%) in catching fraud, the rarity of fraud (only 1% of transactions) means the vast majority of transactions it flags will actually be legitimate. This leads to significant **system design considerations**. Because the occurrence of fraud is rare, so the system must achieve an extremely low false positive rate to be practical. This often means combining multiple detection methods to improve overall reliability. Furthermore, designers must carefully consider the real-world costs of false positives (e.g., blocking legitimate customers and causing frustration) against the costs of false negatives (allowing fraud to occur). At the end of the day, the decision threshold for flagging a transaction cannot be based on accuracy alone but must be adjusted according to specific business requirements and risk tolerance. The **real-world implication** of our calculations is that, with the given parameters, approximately 68% of all flagged transactions will be false alarms. Such a high rate of incorrect alerts is likely to be unacceptable in practice as it would overwhelm investigators, erode customer trust, and diminish the system's operational value.

(d) Statistical Independence Check

Step 1: Definition of statistical independence Two events F and A are statistically independent if:

$$P(F \cap A) = P(F) \cdot P(A)$$

Or equivalently:

$$P(F|A) = P(F) \quad \text{or} \quad P(A|F) = P(A)$$

Step 2: Check independence using calculated values We have:

$$P(F) = 0.01$$

$$P(F|A) \approx 0.3242$$

Since $P(F|A) \neq P(F)$ $(0.3242 \neq 0.01)$, the events are **not independent**.
Step 3: Alternative check using joint probability

$$P(F \cap A) = P(A|F) \cdot P(F) = 0.95 \times 0.01 = 0.0095$$

$$P(F) \cdot P(A) = 0.01 \times 0.0293 = 0.000293$$

Since $P(F \cap A) \neq P(F) \cdot P(A)$ $(0.0095 \neq 0.000293)$, the events are **not independent**.
Step 4: Interpretation The system's flagging behavior is dependent on whether a transaction is fraudulent. This is desirable for a fraud detection system since we want the flagging probability to be different for fraudulent vs legitimate transactions.

$$\boxed{\text{Events F and A are NOT independent}}$$

Question 4: Random Variables

The training time T (in minutes) for a neural network on a random dataset follows this probability distribution:

t (minutes)	10	20	30	40	50
P(T = t)	0.1	0.3	0.4	0.15	0.05

(a) Calculate the expected training time E[T] and the variance Var(T)

Step 1: Set up the probability distribution
We are given the discrete probability distribution for the training time T:

t (minutes)	10	20	30	40	50
P(T = t)	0.1	0.3	0.4	0.15	0.05

Step 2: Calculate E[T] - Expected Value
The expected value of a discrete random variable is defined as the weighted average of all possible values, where the weights are the probabilities of each value:

$$E[T] = \sum_{\text{all } t} t \cdot P(T = t)$$

This means we multiply each possible time value by its probability and sum all these products:

$$E[T] = (10 \times 0.1) + (20 \times 0.3) + (30 \times 0.4) + (40 \times 0.15) + (50 \times 0.05)$$

Now calculate each term:

$$\begin{aligned} 10 \times 0.1 &= 1 \\ 20 \times 0.3 &= 6 \\ 30 \times 0.4 &= 12 \\ 40 \times 0.15 &= 6 \\ 50 \times 0.05 &= 2.5 \end{aligned}$$

Sum all terms:

$$E[T] = 1 + 6 + 12 + 6 + 2.5 = 27.5$$

$$\boxed{E[T] = 27.5 \text{ minutes}}$$

Step 3: Calculate E[T^2] for variance
To calculate the variance, we need E[T^2] first. The formula for E[T^2] is similar to E[T], but we square each t value before multiplying by its probability:

$$E[T] = \sum_{\text{all } t} t \cdot P(T = t)$$

Calculate each squared value:

$$\begin{aligned} 10 &= 100 \\ 20 &= 400 \\ 30 &= 900 \\ 40 &= 1600 \\ 50 &= 2500 \end{aligned}$$

Now we multiply each squared value by its probability:

$$E[T] = (100 \times 0.1) + (400 \times 0.3) + (900 \times 0.4) + (1600 \times 0.15) + (2500 \times 0.05)$$

Calculate each term:

$$\begin{aligned} 100 \times 0.1 &= 10 \\ 400 \times 0.3 &= 120 \\ 900 \times 0.4 &= 360 \\ 1600 \times 0.15 &= 240 \\ 2500 \times 0.05 &= 125 \end{aligned}$$

Sum all terms:

$$E[T] = 10 + 120 + 360 + 240 + 125 = 855$$

Step 4: Calculate Var(T)
The variance measures how spread out the distribution is. The formula for variance is:

$$\text{Var}(T) = E[T] - (E[T])$$

We substitute the values we calculated:

$$\text{Var}(T) = 855 - (27.5)$$

Calculate (27.5):

$$27.5 \times 27.5 = 756.25$$

Now subtract:

$$\text{Var}(T) = 855 - 756.25 = 98.75$$

$$\boxed{\text{Var}(T) = 98.75}$$

(b) The cost of training in dollars is given by C = 50 + 2T

(i) Expected cost E[C]

Step 1: Write the expectation expression

$$E[C] = E[50 + 2T]$$

Step 2: Apply the linearity property of expectation The linearity property states that for any random variable X and constants a and b:

$$E[aX + b] = a \cdot E[X] + b$$

This works because:

- The expectation of a constant is the constant itself: $E[b] = b$
- Constants can be factored out of expectations: $E[aX] = a \cdot E[X]$
- Expectation is linear: $E[X + Y] = E[X] + E[Y]$

Applying this to this then becomes:

$$E[C] = E[50 + 2T] = E[50] + E[2T]$$

Step 3: Calculate each term

$$\begin{aligned} E[50] &= 50 \quad \text{(expectation of a constant is the constant itself)} \\ E[2T] &= 2 \cdot E[T] = 2 \times 27.5 = 55 \quad \text{(constant factors out of expectation)} \end{aligned}$$

Step 4: Combine the results

$$E[C] = 50 + 55 = 105$$

$$\boxed{E[C] = 105 \text{ dollars}}$$

(ii) Variance of cost Var(C)
Step 1: Writing out the variance expression

$$\text{Var}(C) = \text{Var}(50 + 2T)$$

Step 2: Apply the variance property for constants
The variance property states that for any random variable X and constants a and b:

$$\text{Var}(aX + b) = a \cdot \text{Var}(X)$$

This works because:

- Adding a constant to a random variable shifts the distribution but doesn't change its spread, so $\text{Var}(X + b) = \text{Var}(X)$
- Multiplying by a constant scales both the values and the spread, so $\text{Var}(aX) = a \cdot \text{Var}(X)$

Applying this to the problem then becomes:

$$\text{Var}(C) = \text{Var}(50 + 2T) = \text{Var}(2T)$$

Step 3: Apply the scaling property

$$\text{Var}(2T) = 2 \cdot \text{Var}(T) = 4 \times \text{Var}(T)$$

Step 4: Substitute the known variance

$$\text{Var}(C) = 4 \times 98.75 = 395$$

$$\boxed{\text{Var}(C) = 395}$$

(iii) Properties used

For expectation:

- **Linearity of expectation:** $E[aX + b] = a \cdot E[X] + b$ for any constants a, b and random variable X
- **Expectation of constant:** $E[c] = c$ for any constant c

For variance:

- **Variance with constants:** $\text{Var}(aX+b) = a \cdot \text{Var}(X)$ for any constants a, b and random variable X
- **Variance of constant:** $\text{Var}(c) = 0$ for any constant c

(c) The model needs to be trained on 3 independent datasets

(i) Expected total training time E[Ttotal]

Step 1: Write the expectation expression

$$E[T_{\text{total}}] = E[T + T + T]$$

Step 2: Apply linearity of expectation

The linearity property works regardless of whether the variables are independent or not:

$$E[T + T + T] = E[T] + E[T] + E[T]$$

Step 3: Use identical distributions

Since each T has the same distribution as T:

$$E[T] = E[T] = E[T] = E[T] = 27.5$$

Step 4: Calculate the sum

$$E[T_{\text{total}}] = 27.5 + 27.5 + 27.5 = 3 \times 27.5 = 82.5$$

$$\boxed{E[T_{\text{total}}] = 82.5 \text{ minutes}}$$

(ii) Variance Var(Ttotal)

Step 1: Write the variance expression

$$\text{Var}(T_{\text{total}}) = \text{Var}(T + T + T)$$

Step 2: Apply variance of sum for independent variables

For independent random variables, the variance of the sum equals the sum of the variances:

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) \quad \text{when X and Y are independent}$$

This extends to three variables:

$$\text{Var}(T + T + T) = \text{Var}(T) + \text{Var}(T) + \text{Var}(T)$$

Step 3: Use identical distributions Since each T has the same distribution as T:

$$\text{Var}(T) = \text{Var}(T) = \text{Var}(T) = \text{Var}(T) = 98.75$$

Step 4: Calculate the sum

$$\text{Var}(T_{\text{total}}) = 98.75 + 98.75 + 98.75 = 3 \times 98.75 = 296.25$$

$$\boxed{\text{Var}(T_{\text{total}}) = 296.25}$$

(iii) Independence assumption discussion:
The assumption that the training times are independent is reasonable because each dataset is different and randomly chosen. The time it takes to train on one dataset doesn't affect the time it takes to train on another, since each has its own unique size and complexity. In a good experiment, we make sure the datasets are separate so that their training times are unrelated.

However, in the real world, there are several reasons why this might not be true. For example, if all the training runs use the same computer, the first one might make the computer hot, slowing down the next ones. Or, the first training might store some data in memory, making the later trainings faster. Sometimes, the software itself might improve after the first training, changing how long the others take. Also, if the datasets are similar or come from the same source, their training times could be linked. Things like other programs running on the computer or internet traffic could also affect all the training sessions together.

So even though it's usually safe to assume the training times are independent in a well-planned experiment, real-life issues can sometimes make them connected, which would change our calculations.