

CSDS 440: Machine Learning

Soumya Ray (he/him, sray@case.edu)

Olin 516

Office hours T, Th 11:15-11:45 or by appointment

Announcements

- Quiz 1 Thursday, in class, 30 minutes, closed book/notes
 - Topics: everything up to and including decision trees
 - Remember to review probability and statistics

Recap

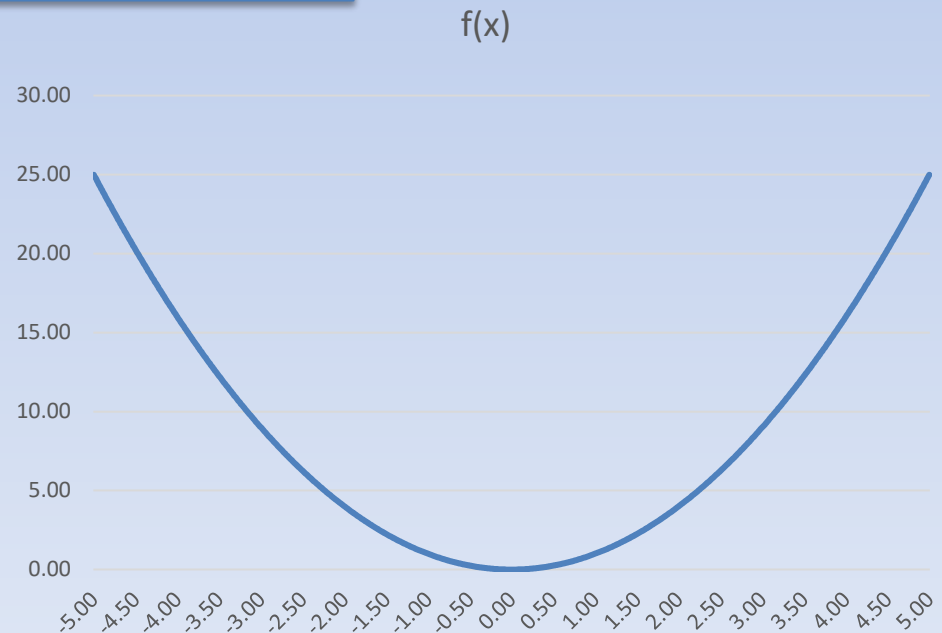
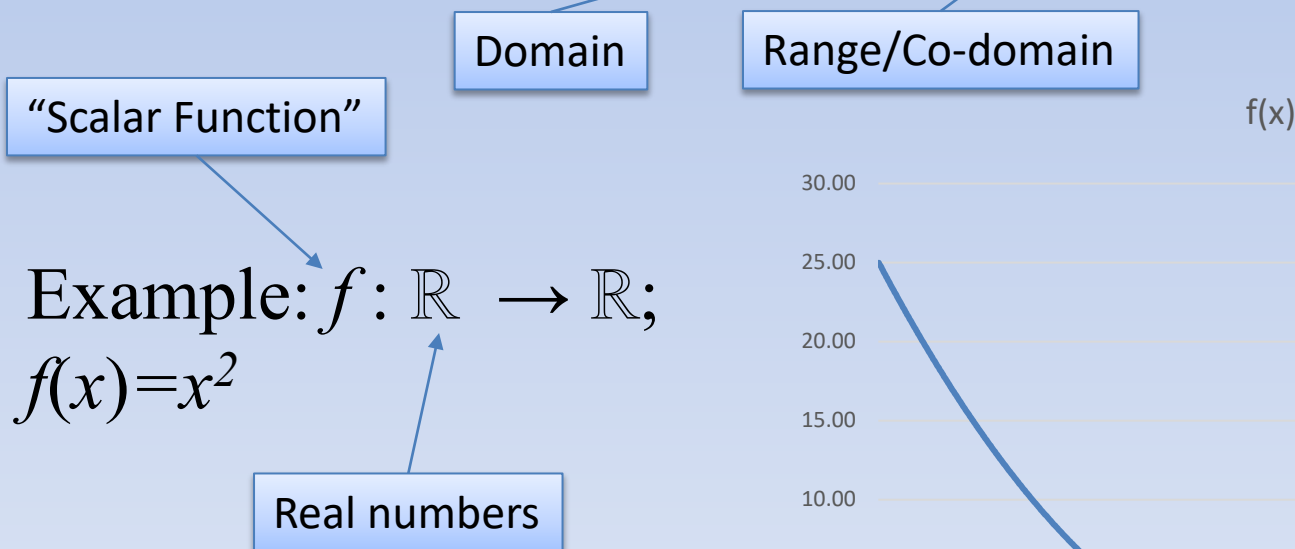
- The c____ t____ tracks four kinds of results the t____ p____, t____ n____, f____ p____ and f____ n____.
- What is accuracy?
- What are the weaknesses of accuracy as a metric? (i) (ii) (iii)
- What is weighted accuracy? Why is it used?
- What is precision? Recall? Why are these used?
- What is specificity? Sensitivity? The F1 score?
- What is a learning curve? Why is this useful?
- When the algorithm produces a confidence, we can draw ____ - ____ graphs or ____ ____ graphs.
- To do this we move a t____ down the list of c____. At each point, examples with confidence h____ than the t____ are classified p____, and the rest classified n____. Then we compute the metrics at this threshold.
- ROC graphs are m____ i____. Also, random classifiers have ROC graphs that are d____.
- We often use a ____ u____ r____ as a comparison statistic. However, ROC graphs can be misleading if the ____.

Review of Calculus and Optimization

- Calculus classes/CSDS 477 / MATH 427/ MATH 433 for the less-crashy version

Functions

- A function *maps* input(s) into output(s)
- Usually denoted $f: X \rightarrow Y; f(x)=y$



Multivariate Functions

- A function can have *multiple* inputs, denoted by $f: A \times B \times \dots \rightarrow Y$
- The “ \times ” is the “Cartesian product” or “cross product”: all possible tuples
- Example: $\mathbb{R} \times \mathbb{R}$: all possible pairs of real numbers e.g. $(1, 1.8)$, $(5.98635435, -3.23456)$, $(\pi, \sqrt{2})$ etc.
 - Often abbreviated as \mathbb{R}^2 (\mathbb{R}^D in general)

Example

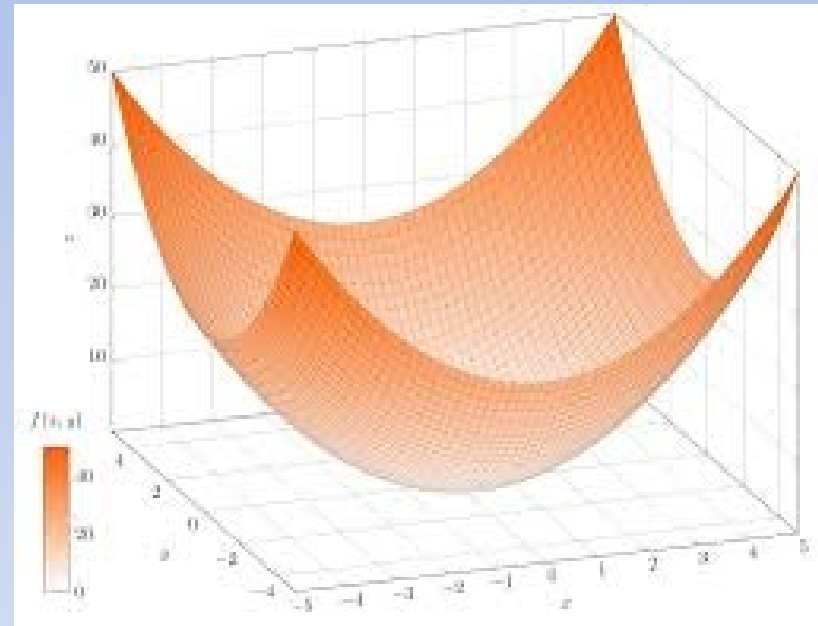
- Suppose we have vectors of size 2
- The norm is a function $\|\cdot\| : \mathbb{R}^2 \rightarrow \mathbb{R}$

Vector functions

- A function can also have multiple *outputs* , denoted by $f: X \rightarrow A \times B \times \dots$
- Example: $f: \mathbb{R} \rightarrow \mathbb{R}^2, f(x) = (x^2, x^3)$
- And a combination of both multiple inputs and outputs $f: A \times B \times \dots \rightarrow C \times D \times \dots$
 - Multivariate vector functions

Gradient of multivariate functions

- When a function takes multiple inputs, we compute *partial derivatives* by varying each input at a time and holding the others fixed
- A function with m inputs will have m partial derivatives



$$f(x, y) = x^2 + y^2$$

Partial Derivatives

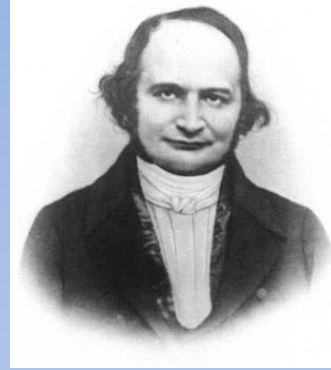
- Suppose $f: \mathbb{R}^2 \rightarrow \mathbb{R}$. The partial derivatives of f are:

$$\left. \frac{\partial f}{\partial x} \right|_{x,y} = \lim_{\Delta x \rightarrow 0} \frac{f(x + \Delta x, y) - f(x, y)}{\Delta x}$$

$$\left. \frac{\partial f}{\partial y} \right|_{x,y} = \lim_{\Delta y \rightarrow 0} \frac{f(x, y + \Delta y) - f(x, y)}{\Delta y}$$

- (Generalizes to functions of n inputs)

Gradient / Jacobian



- The *row* vector

$$\nabla_{x_1, \dots, x_m} f = \left[\frac{\partial f}{\partial x_1} \quad \frac{\partial f}{\partial x_2} \dots \frac{\partial f}{\partial x_m} \right]$$

- Is called the **gradient or Jacobian** of f .

Example

Let $f(x, y) = (x^2 y + xy^3)$. Then

$$\frac{\partial f}{\partial x} = (2xy + y^3), \quad \frac{\partial f}{\partial y} = (x^2 + 3xy^2)$$

$$\nabla f = \begin{bmatrix} 2xy + y^3 & x^2 + 3xy^2 \end{bmatrix}$$

Vector functions

- A vector function $f: \mathbb{R} \rightarrow \mathbb{R}^n$ can be viewed as a *vector of scalar* functions
- Let $f(x) = (x^2, x^3)$; then $f(x) = (f_1(x), f_2(x))$; $f_1(x) = x^2$; $f_2(x) = x^3$

Jacobian of vector functions

- The Jacobian is then the *column* vector:

$$\nabla_x f = \begin{bmatrix} \frac{df_1}{dx} \\ \frac{df_2}{dx} \\ \vdots \\ \frac{df_n}{dx} \end{bmatrix}$$

$$f(x) = (x^2, x^3); \nabla_x f = \begin{bmatrix} 2x \\ 3x^2 \end{bmatrix}$$

Jacobian of Multivariate Vector functions

- Suppose $f: \mathbb{R}^m \rightarrow \mathbb{R}^n$, then

$$\nabla_x f = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \dots & \frac{\partial f_1}{\partial x_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \dots & \frac{\partial f_n}{\partial x_m} \end{bmatrix}$$

Example

- Let $f(x, y) = (x^2y, xy^3)$.

$$\nabla_{x,y} f = \begin{bmatrix} 2xy & x^2 \\ y^3 & 3xy^2 \end{bmatrix}$$

Partial Derivative Shortcuts

$$\frac{\partial}{\partial x}(f(x) + g(x)) = \frac{\partial f(x)}{\partial x} + \frac{\partial g(x)}{\partial x}$$

“Partial Derivative Sum rule”

$$\frac{\partial}{\partial x}(f(x)g(x)) = f(x)\frac{\partial g(x)}{\partial x} + g(x)\frac{\partial f(x)}{\partial x}$$

“Partial Derivative Product rule”

$$\frac{\partial}{\partial x}(f(g(x))) = \frac{\partial f}{\partial g} \frac{\partial g}{\partial x}$$

“Partial Derivative Chain rule”

$$\frac{\partial f}{\partial g} = \frac{\partial f}{\partial x} \bigg|_{g(x)}$$

Note that, in the case of vector and multivariate functions, these are matrix products, so the order is important.

Example

- Let $f(x, y) = (e^x, x+y)$. Let $g(a, b) = (a^2, b^2)$. Then the Jacobian of $f(g(a,b))$ is:

$$\nabla_{x,y} f = \begin{bmatrix} e^x & 0 \\ 1 & 1 \end{bmatrix}; \nabla_{a,b} g = \begin{bmatrix} 2a & 0 \\ 0 & 2b \end{bmatrix}$$

$$\nabla_{a,b} f \circ g = \nabla_g f \nabla_{a,b} g$$

$$\nabla_g f = \nabla_{x,y} f \Big|_{x=a^2, y=b^2} = \begin{bmatrix} e^{a^2} & 0 \\ 1 & 1 \end{bmatrix}$$

$$\nabla_{a,b} f \circ g = \begin{bmatrix} e^{a^2} & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 2a & 0 \\ 0 & 2b \end{bmatrix} = \begin{bmatrix} 2ae^{a^2} & 0 \\ 2a & 2b \end{bmatrix}$$

Higher Order Derivatives

- We can take the derivative of a function multiple times (if possible)
- The rate of change of the derivative is the *second derivative*:

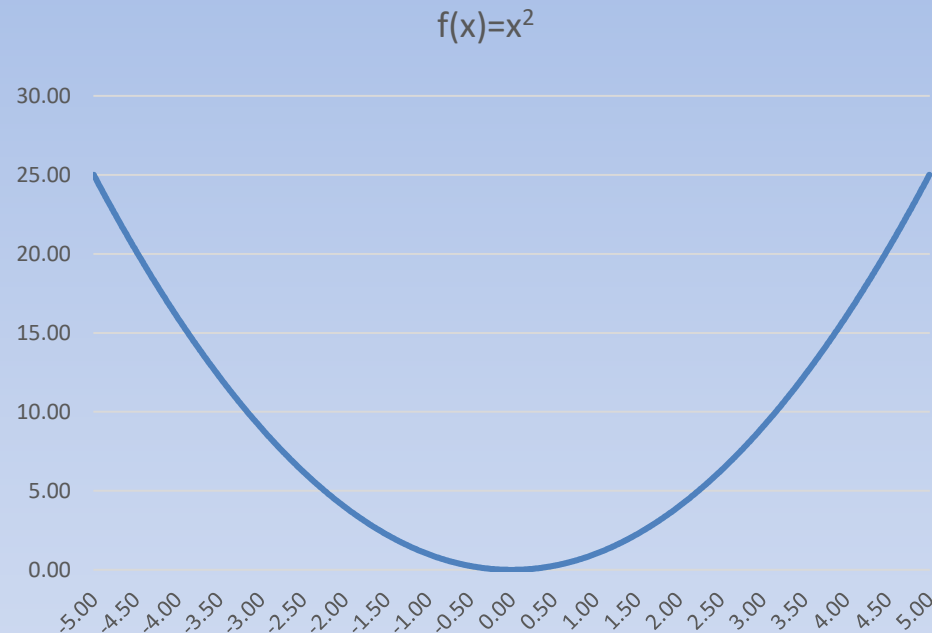
$$\left. \frac{d^2 f}{dx^2} \right|_{x_0} = f''(x_0) = \lim_{\Delta x \rightarrow 0} \frac{f'(x_0 + \Delta x) - f'(x_0)}{\Delta x}$$

Example and Geometry

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

$$\frac{df}{dx} = 2x$$

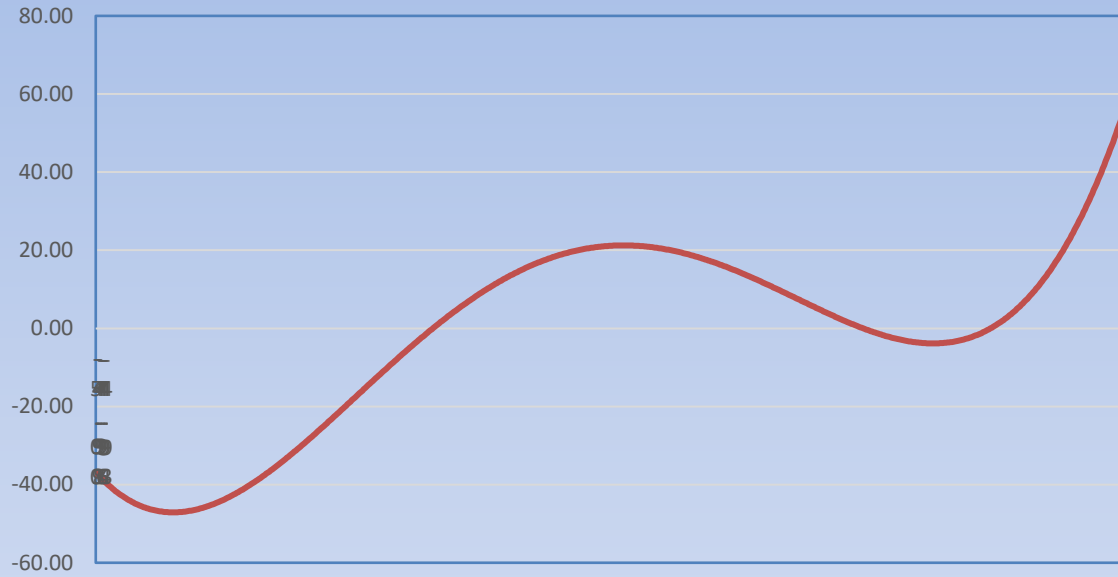
$$\frac{d^2 f}{dx^2} = 2.$$



Function with “positive curvature”

Example and Geometry

$$g(x)=x^4+7x^3+5x^2-17x+3$$



$$g'(x) = 4x^3 + 21x^2 + 10x - 17$$

$$g''(x) = 12x^2 + 42x + 10$$

$$= 64, -25.28, 45.28 \text{ (for different inputs } x\text{)}$$

Multivariate functions

- Suppose we have a function $f: \mathbb{R}^m \rightarrow \mathbb{R}$
- The *derivative is itself a vector function* $\mathbb{R}^m \rightarrow \mathbb{R}^m$:

$$\nabla_{x_1, \dots, x_m} f = \left[\frac{\partial f}{\partial x_1} \quad \frac{\partial f}{\partial x_2} \dots \frac{\partial f}{\partial x_m} \right]$$

- The second derivative takes this as input

Hessian Matrix



- Taking the second derivative creates a $m \times m$ matrix:

$$\nabla_x^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_m \partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_m} & \cdots & \frac{\partial^2 f}{\partial x_m^2} \end{bmatrix}$$

Example

Let $f(x, y) = (x^2 y + xy^3)$. Then

$$\frac{\partial f}{\partial x} = (2xy + y^3), \quad \frac{\partial f}{\partial y} = (x^2 + 3xy^2)$$

$$\nabla_{x,y} f = g(x, y) = \begin{bmatrix} 2xy + y^3 & x^2 + 3xy^2 \end{bmatrix}$$

$$\nabla_{x,y}^2 f = \nabla g_{x,y} = \begin{bmatrix} 2y & 2x + 3y^2 \\ 2x + 3y^2 & 6y \end{bmatrix}$$

Tensors

- What if $f: \mathbb{R}^m \rightarrow \mathbb{R}^n$?

$$\nabla_x f = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \dots & \frac{\partial f_1}{\partial x_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \dots & \frac{\partial f_n}{\partial x_m} \end{bmatrix}$$

Then the second derivative is a **three dimensional** matrix H of size $(n \times m) \times m$ where $H_{ijk} = \frac{\partial f_i}{\partial x_k \partial x_j}$. Such matrices are called **tensors**.

Example

- Let $f(x, y) = (x^2y, xy^3)$.

$$\nabla_{x,y} f = \begin{bmatrix} 2xy & x^2 \\ y^3 & 3xy^2 \end{bmatrix}$$

$$\nabla_{x,y}^2 f = \begin{bmatrix} 2y; 2x & 2x; 0 \\ 0; 3y^2 & 3y^2; 6xy \end{bmatrix}$$

Gradients of Matrices

- Suppose we have a function $f: B \rightarrow A$, $B \in \mathbb{R}^{m \times n} \rightarrow A \in \mathbb{R}^{p \times q}$
- The Jacobian of f will be a four dimensional tensor $\mathbb{R}^{(m \times n) \times (p \times q)}$
- Where $J(i, j, k, l) = \frac{\partial A_{ij}}{\partial B_{kl}}$

Gradient of a vector wrt a matrix

- Suppose we have function $f(A): y=Ax, A \in \mathbb{R}^{m \times n} \rightarrow y \in \mathbb{R}^m$
- What is $J(f)$?

$$\nabla_x f = \begin{bmatrix} \frac{df_1}{dx} \\ \frac{df_2}{dx} \\ \vdots \\ \frac{df_n}{dx} \end{bmatrix} \quad \nabla_A f = \begin{bmatrix} \frac{\partial f_1}{\partial A} \\ \frac{\partial f_2}{\partial A} \\ \vdots \\ \frac{\partial f_m}{\partial A} \end{bmatrix} \quad \frac{\partial f_i}{\partial A} \in \mathbb{R}^{1 \times m \times n}$$

Reshaping Tensors

- In practice, it can often be useful to convert a tensor into a matrix or vector
 - Done by “stacking” tensors
- Idea: Every matrix in $\mathbb{R}^{m \times n}$ can be written as a vector in \mathbb{R}^{mn}
- Similarly, every tensor in $\mathbb{R}^{(m \times n) \times (p \times q)}$ can be written as a matrix in $\mathbb{R}^{mn \times pq}$

What is an optimization problem?

- Find the *extreme values* of a function (called an “**objective function**”)
 - Sometimes we are interested in the extreme values themselves
 - Other times we are interested in the *arguments* to the function that produce those extreme values
 - argmax , argmin

Types of Optimization Problems

- Discrete vs continuous
 - Objective function is defined on discrete or continuous space
- Unconstrained vs constrained
 - Whether there are additional function constraints defining the “feasible region”
- In this class, we will mainly be interested in continuous problems, both unconstrained and constrained
 - Use tools from calculus and linear algebra

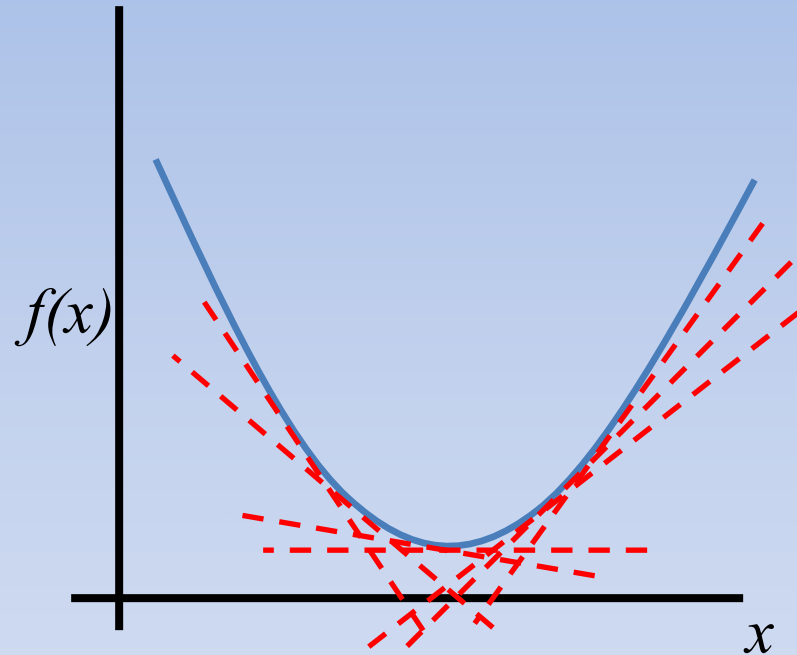
Unconstrained Continuous Optimization

Function of one variable:

$$\min_x f(x)$$

$$\frac{df}{dx} = 0$$

$$\frac{d^2 f}{dx^2} > 0$$



Example



$$g'(x) = 4x^3 + 21x^2 + 10x - 17$$

$$g'(x) = 0 \text{ for } x = -4.5, -1.4, 0.7$$

$$g''(x) = 12x^2 + 42x + 10$$

$$= 64, -25.28, 45.28$$