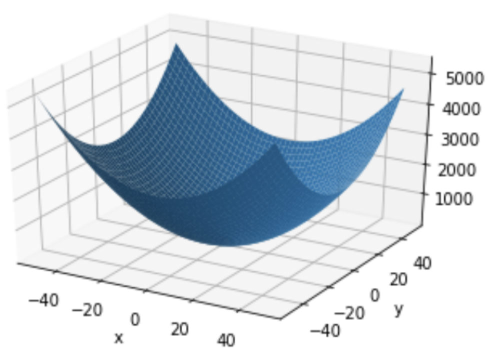
DS 100/200: Principles and Techniques of Data Science **Date: November 6, 2019**

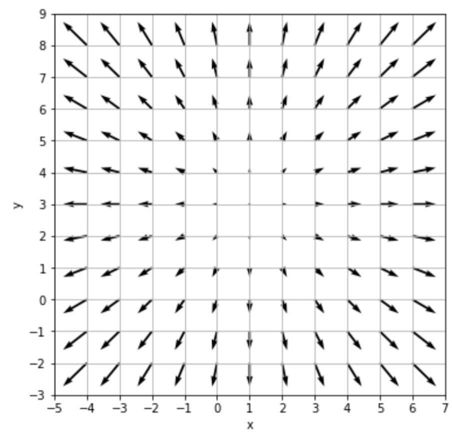
Discussion #11

Name:

Gradients

1. On the left is a 3D plot of $f(x, y) = (x-1)^2 + (y-3)^2$. On the right is a plot of its **gradient field**. Note that the arrows show the relative magnitudes of the gradient vector.





   (a) From the visualization, what do you think is the minimal value of this function and where does it occur?

   (b) Calculate the gradient $\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \end{bmatrix}^T$.

   (c) When $\nabla f = \mathbf{0}$, what are the values of x and y?

Gradient Descent Algorithm

2. Given the following loss function and $\mathbf{x} = (x_i)_{i=1}^{n}$, $\mathbf{y} = (y_i)_{i=1}^{n}$, β^t, explicitly write out the update equation for β^{t+1} in terms of x_i, y_i, β^t, and α, where α is the step size.

$$L(\beta, \mathbf{x}, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n} \left(\beta^2 x_i^2 - log(y_i) \right)$$

Convexity

3. Convexity allows optimization problems to be solved more efficiently and for global optimums to be realized. Mainly, it gives us a nice way to minimize loss (i.e. gradient descent). There are three ways to informally define convexity.

   a. Walking in a straight line between points on the function keeps you above the function. This works for any function.

   b. The tangent line at any point lies below the function (globally). The function must be differentiable.

   c. The second derivative is non-negative everywhere (aka ”concave up” everywhere). The function must be twice differentiable.

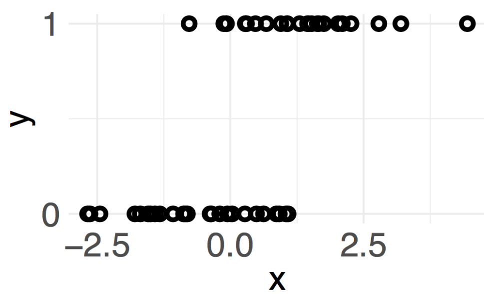
   (a) Is the function described in question 1 convex? Make an argument visually.

   (b) Find a counterexample for the claim that the composition of two convex functions is also convex. $h = g(f(x))$

Logistic Regression

The next two questions refer to a binary classification problem with a single feature x.

4. Based on the scatter plot of the data below, draw a reasonable approximation of the logistic regression probability estimates for $\mathbb{P}(Y = 1 \mid x)$.



5. You have a classification data set consisting of two (x, y) pairs $(1, 0)$ and $(-1, 1)$.

   The covariate vector $\mathbf{x}$ for each pair is a two-element column vector $[1 \quad x]^T$.

   You run an algorithm to fit a model for the probability of $Y = 1$ given $\mathbf{x}$:

$$\mathbb{P}(Y = 1 \mid \mathbf{x}) = \sigma(\mathbf{x}^T \beta)$$

   where

$$\sigma(t) = \frac{1}{1 + \exp(-t)}$$

   Your algorithm returns $\hat{\beta} = \begin{bmatrix} -\frac{1}{2} & -\frac{1}{2} \end{bmatrix}^T$

   (a) Calculate $\hat{\mathbb{P}}\left(Y = 1 \mid \mathbf{x} = [1 \quad 0]^T\right)$

   (b) The empirical risk using log loss (a.k.a., cross-entropy loss) is given by:

$$\begin{aligned} R(\beta) &= \frac{1}{n} \sum_{i=1}^{n} -\log \hat{\mathbb{P}}(Y = y_i \mid \mathbf{x_i}) \\ &= -\frac{1}{n} \sum_{i=1}^{n} y_i \log \hat{\mathbb{P}}(Y = 1 \mid \mathbf{x_i}) + (1 - y_i) \log \hat{\mathbb{P}}(Y = 0 \mid \mathbf{x_i}) \end{aligned}$$

And $\hat{\mathbb{P}}(Y=1 \mid \mathbf{x_i}) = \frac{\exp(\mathbf{x_i}^T\beta)}{1+\exp(\mathbf{x_i}^T\beta)}$ while $\hat{\mathbb{P}}(Y=0 \mid \mathbf{x_i}) = \frac{1}{1+\exp(\mathbf{x_i}^T\beta)}$. Therefore,

$$\begin{aligned} R(\beta) &= -\frac{1}{n}\sum_{i=1}^{n} y_i \log \frac{\exp(\mathbf{x_i}^T\beta)}{1+\exp(\mathbf{x_i}^T\beta)} + (1-y_i)\log\frac{1}{1+\exp(\mathbf{x_i}^T\beta)} \\ &= -\frac{1}{n}\sum_{i=1}^{n} y_i\mathbf{x}_i^T\beta + \log(\sigma(-\mathbf{x}_i^T\beta)) \end{aligned}$$

Let $\beta = [\beta_0 \quad \beta_1]$. Explicitly write out the empirical risk for the data set $(1,0)$ and $(-1,1)$ as a function of β_0 and β_1.

(c) Calculate the empirical risk for $\hat{\beta} = \begin{bmatrix}-\frac{1}{2} & -\frac{1}{2}\end{bmatrix}^T$ and the two observations $(1,0)$ and $(-1,1)$.