INFO251 – Applied Machine Learning

Lab 6

Suraj R. Nair

Announcements

- Quiz 1 on Feb 27
- Lab schedule:
  - Today: Cross Validation, Normalization, Standardization + Gradient Descent Demo
  - Feb 21: Gradient Descent (~20 mins) + Review for Quiz 1 (~30 min)

Topics

- Gradient descent
  - Random initialization, learning rate, iterations, stopping conditions
- Convexity

Gradient Descent

- "A first-order iterative algorithm for finding a **local minimum** of a **differentiable** multivariate function."
  - Stochastic: 1 example per step
  - Batch: all data taken into consideration
  - Mini-batch: use a batch of a fixed number of examples
- Random initialization
- Step size / learning rate
- Stopping conditions (tolerance)

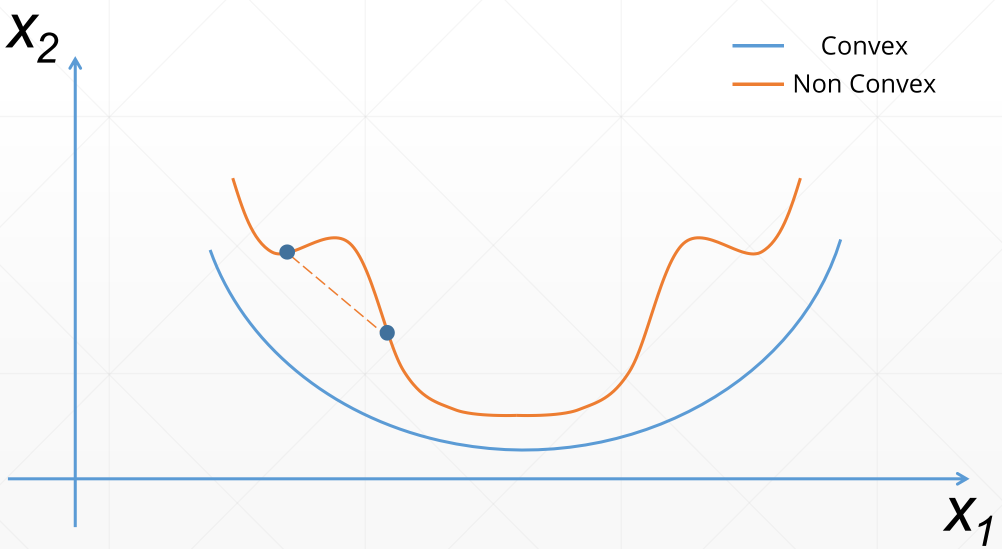
Gradient Descent

1. Begin at a random point
2. Calculate the function value at the point and the gradient (partial derivatives)
3. Pick a new point, move in the direction of steepest descent. The size of the step is governed by the **learning rate**.
4. Repeat!

$$\mathbf{b} = \mathbf{a} - \gamma \, \nabla \, \mathbf{f}(\mathbf{a})$$

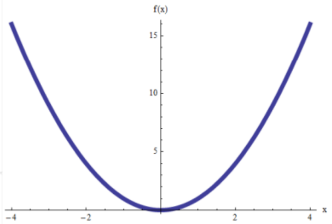
Convexity

- **Convex function:** Second derivative is always nonnegative
- **Graphical interpretation**: Line segment between any two points on the graph of the function does not lie below the graph between the two points

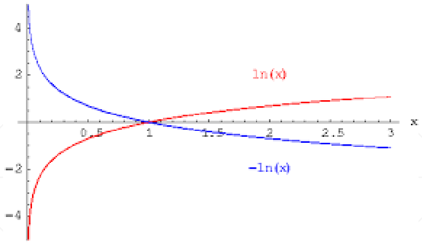


Convexity

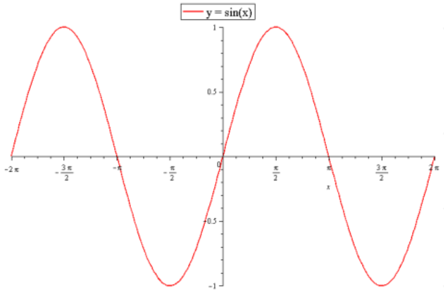
$f(x) = x^2$



$f(x) = \ln(x)$

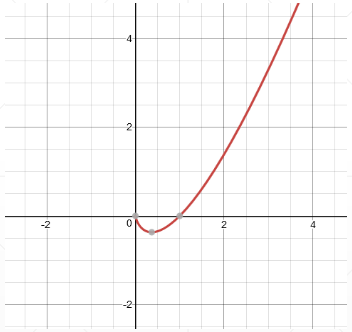


$f(x) = \sin(x)$

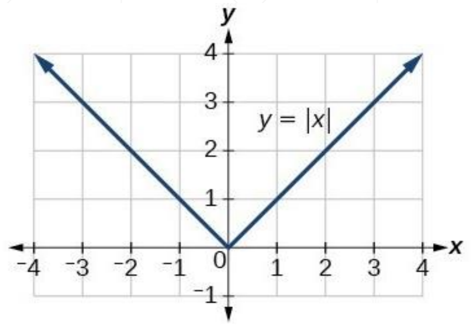


$f(x) = x^3$

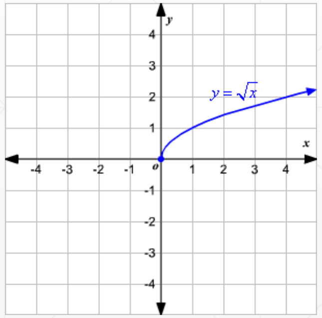
$f(x) = x\ln(x)$



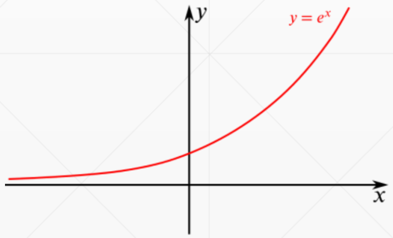
$f(x) = |x|$



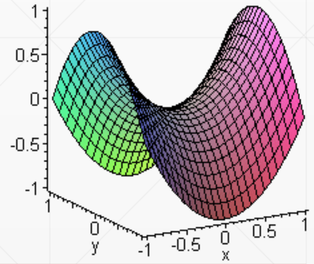
$f(x) = x^{1/2}$



$f(x) = e^x$



$f(x) = ax^2 - by^2$



Quiz 1 Review

Regression

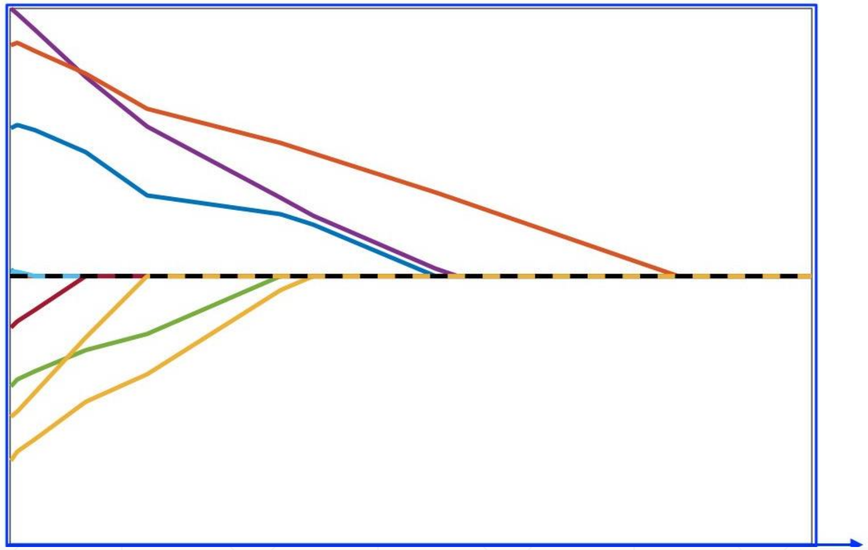
- When comparing two regression models, the model that produces the higher R2 will provide less biased estimates of the causal impact of the independent variables on the dependent variable:
  - True
  - False

Difference-in-difference

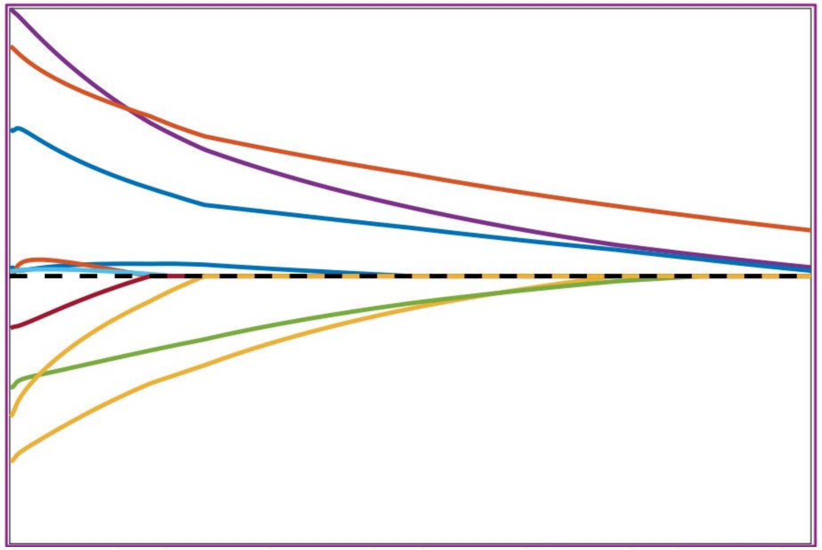
- The key identifying assumption is
  A. Outcomes in the control and treatment group would have been the same in the absence of treatment
  B. Trends in the control and treatment group would have been the same in the absence of treatment
  C. Outcomes pre- and post-treatment would have been the same in the absence of treatment
  D. Outcomes pre-treatment would have been the same in the absence of treatment

Regularization

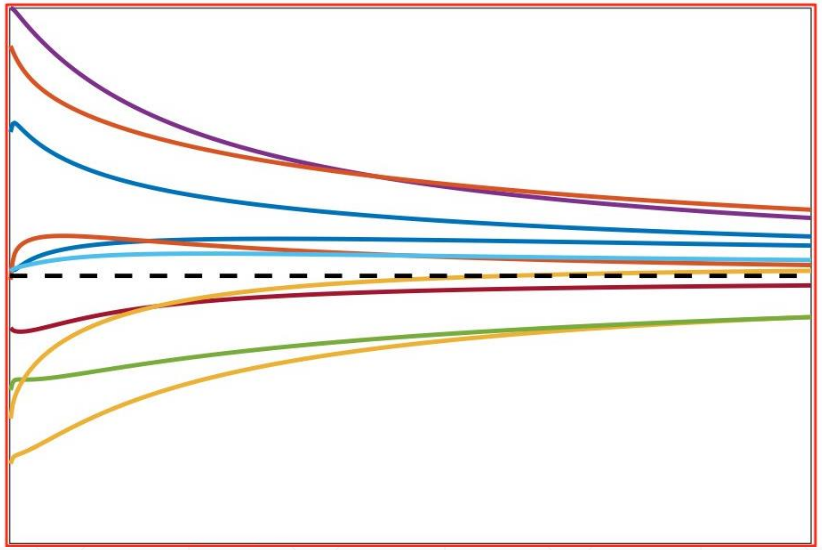
- Match the penalty (Lasso, Ridge, ElasticNet) to the coefficient plot



A



B



C

Decision Boundaries

- Which of the following algorithms recovers non-linear decision boundaries:
  - K-nearest neighbors (K = 5)
  - Linear Regression
  - Logistic Regression
  - Linear Regression with lasso regularization

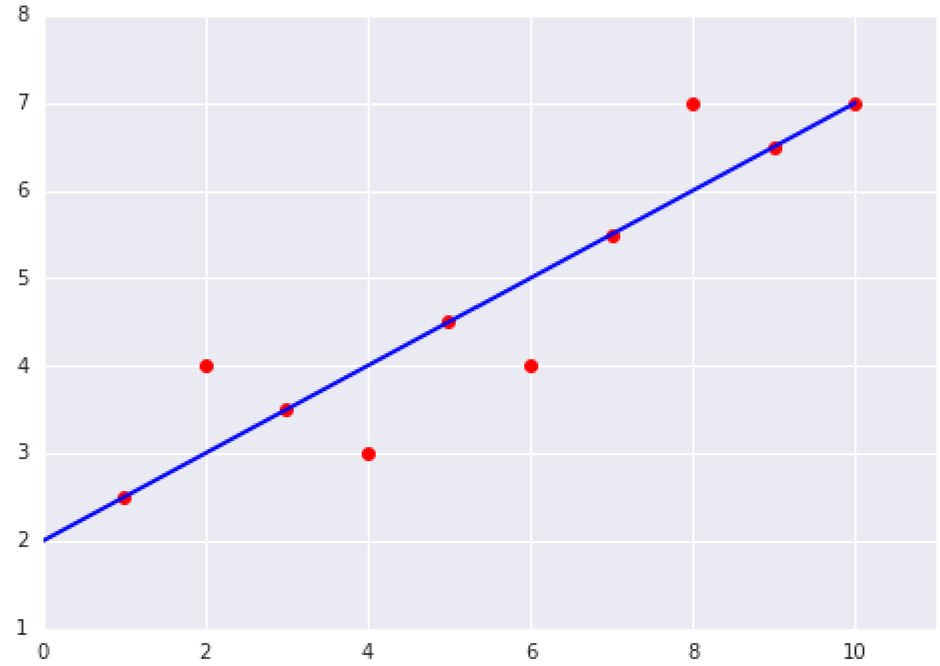
Gradient Descent

- You are trying to find the parameters for a multivariate linear regression using gradient descent. The algorithm is initialized at some random starting point. However, it is taking very long to converge (> 10,000 iterations). What could be the reason(s)?
  - Step size is too small
  - Step size is too large
  - Data may not have been scaled

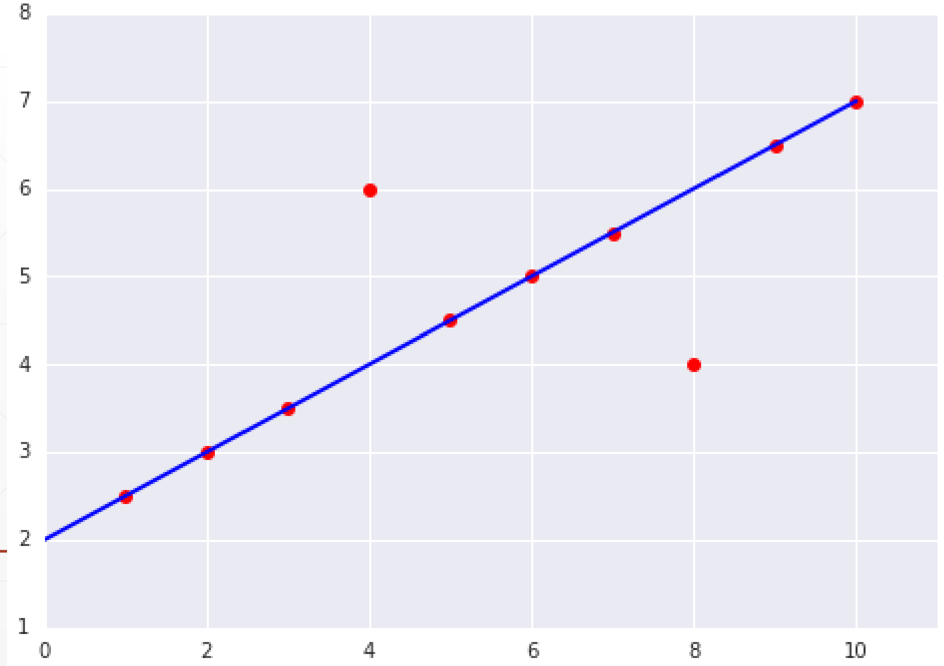
Mean Squared Error

- Suppose you build a linear regression model which predicts y = f(x). Which of these two cases has a higher MSE?
  - A
  - B

A



B



Logistic Regression

- Example with single predictor variable
  - Likelihood of honor student, by major
    - $\text{logit}(honor_i) = \alpha + \beta STEM_i + \epsilon_i$
  - exp(0.593) = 1.809
    - (this is the odds ratio)
    - (corresponds to p=0.644)
  - The odds radio can also be seen in the cross-tabs:
    - Odds for non-STEM: 0.23 (17/74)
    - Odds for STEM: 0.42 (32/77)
    - Odds for STEM 81% higher
      - 0.42 / 0.23 = 1.809
      - 0.644 / (1-0.644) = 1.809

Logistic regression

Number of obs	=	200
LR chi2(1)	=	3.10
Prob > chi2	=	0.0781
Pseudo R2	=	0.0139

Log likelihood = -109.80312

hon	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
stem	.5927822	.3414294	1.74	0.083	-.0764072	1.261972
intercept	-1.470852	.2689555	-5.47	0.000	-1.997995	-.9437087

	stem		
hon	**no**	**yes**	Total
0	74	77	151
1	17	32	49
Total	91	109	200