1. Two major ways to do multinomial eval:
    a. Softmax Loss
    b. One vs. All with binary (logistic) function
2. Naming -
    a. “Logistic” regression due to Sigmoid (logistic) function
    b. “Softmax” regression due to softmax function
3. No closed form solution, despite convexity
4. Many, many optimizers:
    a. Newton / Newton-CG
    b. BFGS
        i. L-BFGS
    c. IRLS
    d. Trust Region Conjugate Gradient
    e. Gradient Descent
        i. GD + Line Search
    f. Stochastic Average Gradient
5. Difficult Bayesian Solutions (No convenient conjugate prior)
6. Discriminative (Learns P(Y|X), rather than first the joint P(Y, X) and then conditioning on X (the generative approach))
7. Without regularization, the weights will become arbitrary large, damaging generalization. Penalties are more important than in the regression setting.
8. You can get better generalization with a stochastic solver [https://arxiv.org/pdf/1708.05070.pdf]
9. The reason scaling can still be important is for the optimizer - even though you technically have a convex model and will get the same solution
10. Linear generalization is stronger than almost every other form of generalization for unstructured data (trees + networks overfit)
11. Every relationship between your feature and the label should be as close to linear as possible
12. You can use boxcox transform to automatically get close to linear