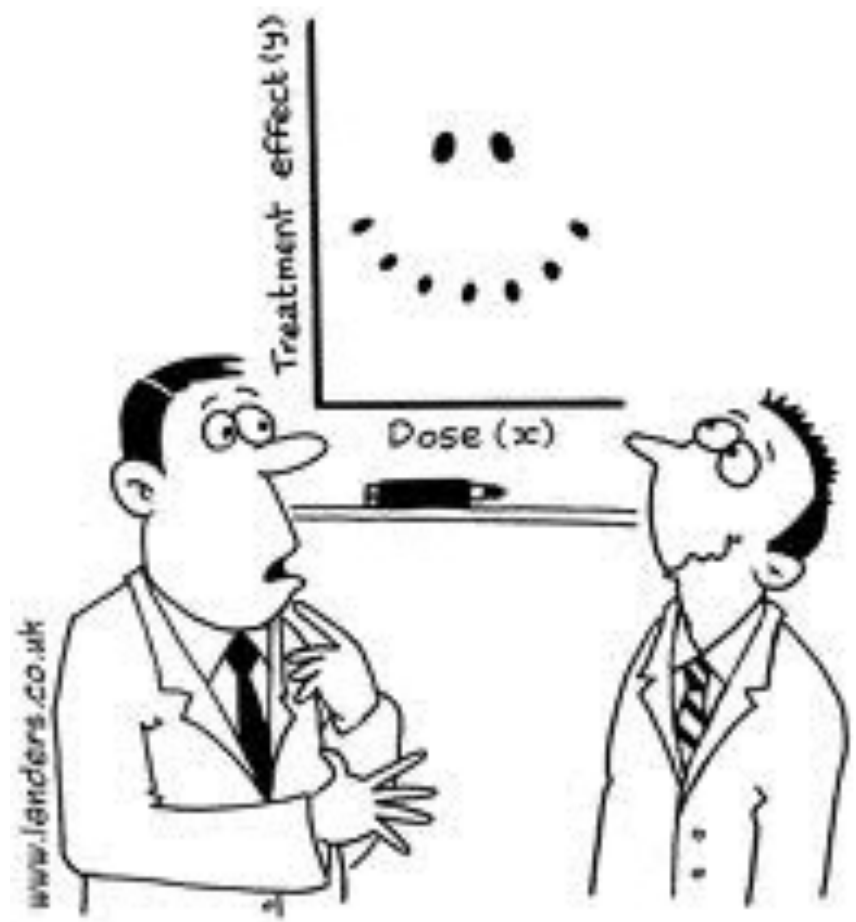
Linear regression (continued…)



"It's a non-linear pattern with outliers.....but for some reason I'm very happy with the data."

Reminders

- Assignment #2 is released
  - some implementation questions for practical regression
  - a strong focus on calculus and derivatives
- Thought questions due next week

Thought question

- In Maximum Likelihood Estimation, should the probability function be always convex? If yes, how to deal with non convex functions?
  - The negative log of the likelihood and prior may not be convex, depending on the choice —> many cases it is not convex
  - Might explicitly choose likelihoods and priors to ensure convexity
    - called log-concave function
  - There are techniques to find the minima of non-convex functions; generally, only local solutions are found (not global solutions)
  - The field called "global optimization" tries to guarantee global solutions to non-convex problems
  - One common solution: random restarts, keep best found solution

Thought question

- Can independent variables be looked at as features which don't change in relation to another feature? If so, then why are independent variables important in machine learning?
  - Even if the features do not change in relation to each other, they may still change in relation to a desired (target) variable
  - If **all features and targets** were independent random variables, then learning a prediction function using the features would not be useful
  - Having independent features that are correlated with a separate target variable can make learning simpler, since these features more clearly contribute to changes in the target

Maximum likelihood

- Assume that there is noise in the measurement of the target
  - but no noise in the measurement of X
  - the noise in measuring y is independent of x
- Then maximum likelihood parameter w are given by the ordinary least-squares solution
- Now we need to examine
  - extensions to multiple targets
  - properties of the solution (including variance)
  - practicality and feasibility of this optimization in real-world scenarios

Recall

- We re-wrote the maximum likelihood optimization as a minimization with matrix and vector variables X, y and w

- Then took the gradient w.r.t. to vector w and solved for w

- Then checked the Hessian at that solution, and found it was positive semi-definite, so the solution is a local minimum

    - because Hessian positive semi-definite for all w, this solution is actually a global minimum, since this indicates the loss is convex

    - could also have first checked if the loss was convex; if so, then the found minimum is a global minimum

- Simple optimization skills important, as most ML algorithms based on minimizing (or maximizing) objectives

More intuition on solution

$$\mathbf{w}^* = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \qquad \hat{\mathbf{y}} = \mathbf{X}\mathbf{w}^*$$

- Gradient being zero gives a stationary point
  - only in a few cases can we solve the equation gradient E(w) = 0
  - e.g., we will not be able to do so for logistic regression
  - for other cases we will step in the direction of the gradient until we reach such a stationary point
- Hessian (locally) tells you how the gradient changes
  - can write the problem in terms of directional derivatives
  - then get a condition that reduces to univariate derivatives

Directional second derivative

$$\text{At stationary point } \mathbf{w}^*, \nabla f(\mathbf{w}) = \mathbf{0}$$

$$\mathbf{w}(t) = \mathbf{w}^* + t\mathbf{w}$$

$$g(t) = f(\mathbf{w}(t))$$

$$g'(0) = \nabla f(\mathbf{w}(t))^\top \mathbf{w} = 0$$

$$g''(0) = \mathbf{w}^\top \nabla^2 f(\mathbf{w}(t))^\top \mathbf{w}$$

Intuition for second derivative test in univariate setting

$$0 < f''(x) = \lim_{h \to 0} \frac{f'(x+h) - f'(x)}{h} = \lim_{h \to 0} \frac{f'(x+h) - 0}{h} = \lim_{h \to 0} \frac{f'(x+h)}{h}.$$

Thus, for h sufficiently small we get

$$\frac{f'(x+h)}{h} > 0$$

Exercise: positive definite and positive semi-definite

- Recall that $H = 2X^T X$
- H is positive semi-definite if $z^T H z \geq 0$ for all $z \neq 0$
- H is positive definite if $z^T H z > 0$ for all $z \neq 0$
- Why is H positive definite if X has linearly independent columns?
- Why is H positive semi-definite if X has linearly dependent columns?
- Multiple ways to see this, using definition of linearly dependent vectors and eigenvalue decomposition.

Example: OLS

Example 11: Consider again data set $\mathcal{D} = \{(1, 1.2), (2, 2.3), (3, 2.3), (4, 3.3)\}$

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}, \quad \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 1.2 \\ 2.3 \\ 2.3 \\ 3.3 \end{bmatrix},$$

In Matlab, can compute

1. $\mathbf{X}^\top \mathbf{X}$
2. $(\mathbf{X}^\top \mathbf{X})^{-1}$
3. $(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$

$$\mathbf{X}^\top \mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}$$

What if we did not add the column of 1s?

Whiteboard

- Weighted error functions, if certain data points “matter” more than others
- Predicting multiple outputs (multivariate y)
- Expectation and variance for the solution vector

Linear regression for non-linear problems

$$f(x) = w_0 + w_1 x, \longrightarrow f(x) = \sum_{j=0}^{p} w_j x^j,$$

$$\begin{array}{cc} & \mathbf{X} \\ 1 & \begin{array}{|c|}\hline x_1 \\ \hline x_2 \\ \hline \dots \\ \hline x_n \\ \hline \end{array} \\ n & \end{array} \rightarrow \begin{array}{c} \mathbf{\Phi} \\ \begin{array}{|c|c|c|}\hline \phi_0(x_1) & \dots & \phi_p(x_1) \\ \hline \dots & \dots & \dots \\ \hline \dots & \dots & \dots \\ \hline \phi_0(x_n) & \dots & \phi_p(x_n) \\ \hline \end{array} \end{array}$$

Figure 4.3: Transformation of an $n \times 1$ data matrix $\mathbf{X}$ into an $n \times (p+1)$ matrix $\mathbf{\Phi}$ using a set of basis functions ϕ_j, $j = 0, 1, \dots, p$.

$$\mathbf{w}^* = \left(\mathbf{\Phi}^\top \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^\top \mathbf{y}.$$

Overfitting

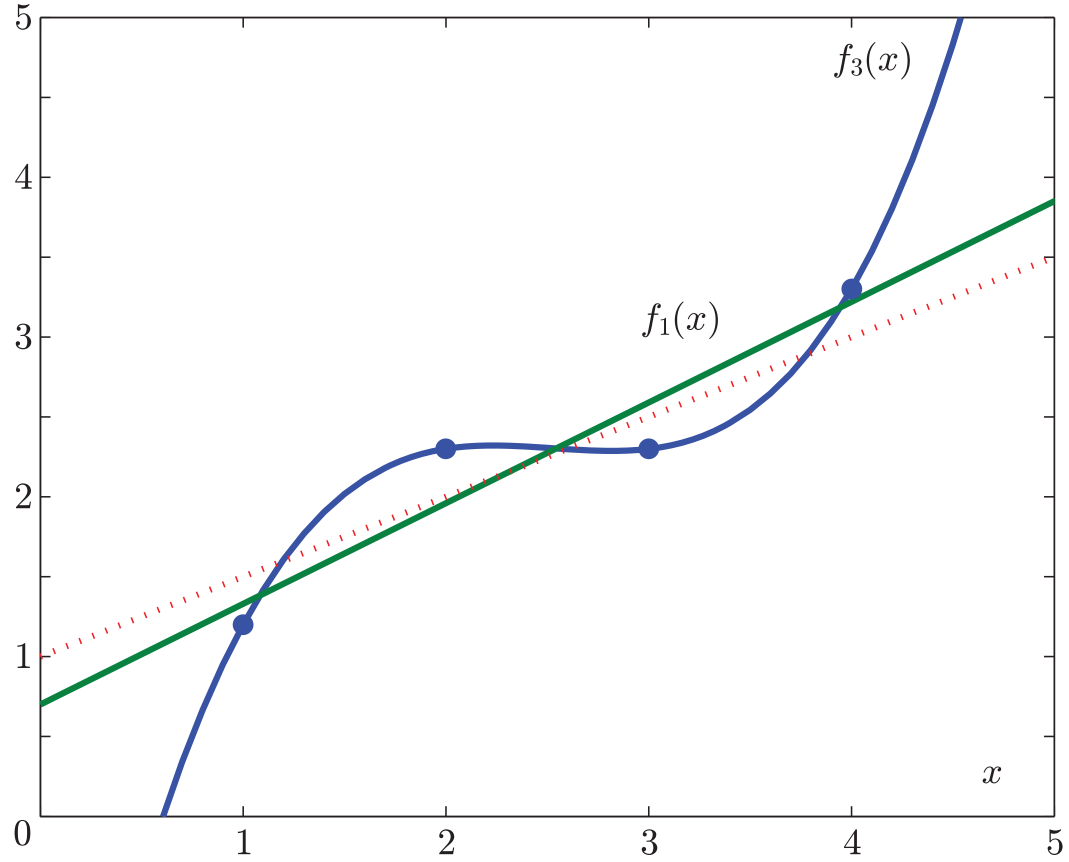


Figure 4.4: Example of a linear vs. polynomial fit on a data set shown in Figure 4.1. The linear fit, $f_1(x)$, is shown as a solid green line, whereas the cubic polynomial fit, $f_3(x)$, is shown as a solid blue line. The dotted red line indicates the target linear concept.

$$\mathbf{w}_1^* = (0.7, 0.63)$$

$$\mathbf{w}_3^* = (-3.1, 6.6, -2.65, 0.35)$$