SMGT 432: Validation & Regularization Lecture

Announcements

1. Monday will feature guest speakers from Rice Soccer
2. Reading #2 has been posted (due Wednesday)
3. What IDEs did you all use last Pyday?

Bring up the questions from last lecture.

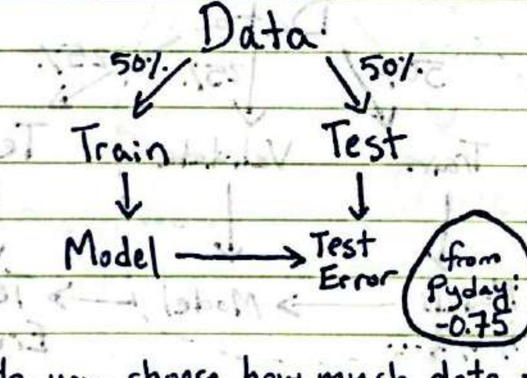
Questions

1. What is the benefit of regularization in a regression model?
2. What are the tradeoffs when choosing λ in ridge regression?
3. What are the tradeoffs when choosing K in K-fold cross validation?

Pause What if we just calculate $\frac{1}{n}\sum_{i=1}^{n} \log(\hat{p}_i^{y_i})$?

from Pyday: -0.68

Validation



Pause How do you choose how much data to hold out for the test set? What if you hold out too little? Too much?

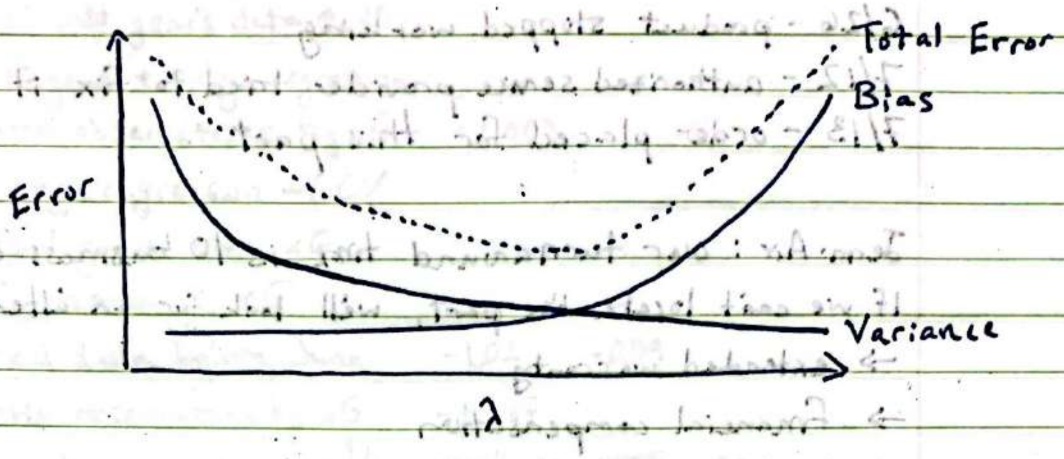
Regularization

$$\hat{\alpha}, \hat{\beta} = \arg\min_{\alpha, \beta} -\ell(\alpha, \beta; X, y) + \lambda \sum_{t=1}^{T} \beta_t^2$$

Why on Earth would you do this?

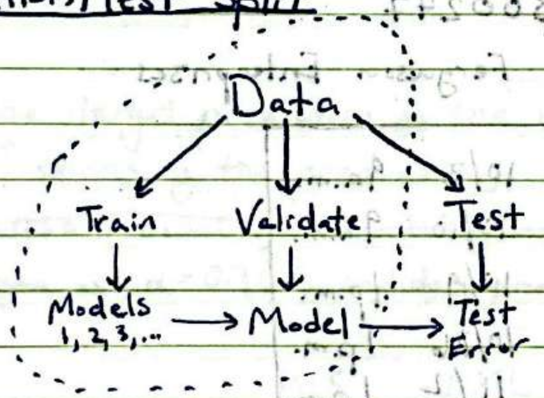
Pause What is λ? What if it's zero? Infinity?

Bias-Variance Tradeoff



Pause | How do we find the right value for λ?
Answer: Validation! But then what's the test error?

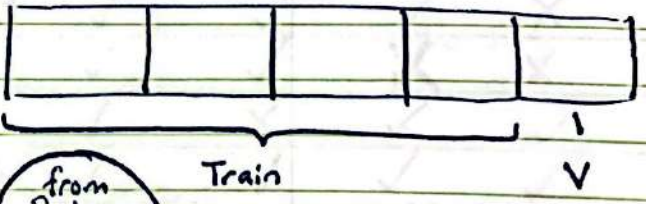
Train/Validation/Test Split



Pause | Now how much data do we allocate to ~~Train~~/Validate/Test?

Cross-Validation

1. Split data into K "folds" (e.g. $K = 5$)
2. For each fold $k = 1, \ldots, K$:
   a. Train model WITHOUT fold k
   b. Produce predictions for fold k
   c. Evaluate error on fold k
3. Calculate average error across folds



from Pyday
$\lambda = 0.0045$
-0.74

Pause | What's the smallest K could be? Largest? What's the tradeoff?