Controlling Model Complexity

Dale Smith

Atlanta, GA

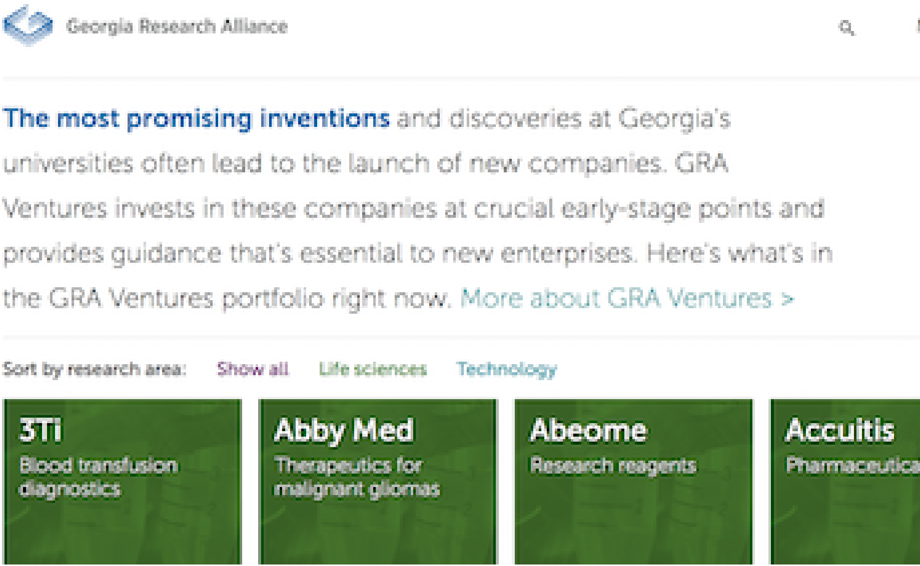
January 18, 2018

Overview

BioTech - Opportunities of Scale

- Cost of sequencing a single person's genome - \$3 bln 1989 – 2001 to under \$1k today
- Twelve years to sequence one versus multiple person's genome in under a week
- Sequence DNA from skin cancer tumor versus sequence DNA from skin cells
- Individualized treatments
- Labiotech.eu: "The Robots are Coming: Is AI the Future of Biotech?"

Biotech Firms in Atlanta



Models are Algorithms Fit to Data

$$y = f(x) \tag{1}$$

- x – called features or independent variables
- y – response variable, predicted quantity, dependent variable dataset

Models are Algorithms Fit to Data

$$y = f(x) \tag{2}$$

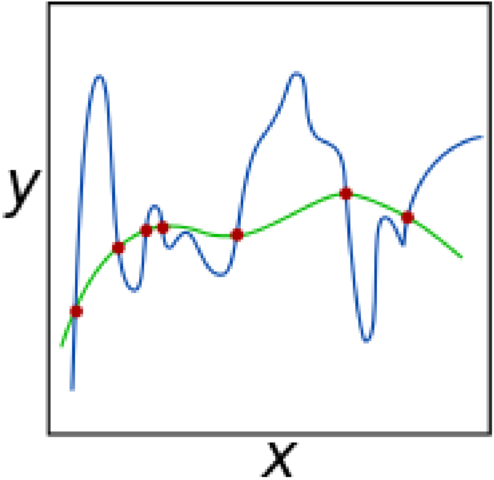
- Use future x data in the model to generate a y
- If y is a set of classes, the problem is called a classification problem
- Otherwise, y is a continuous variable and it's a regression problem
- Separate dataset into training, (validation), test sets, reflecting the characteristics of the undivided dataset

Model Complexity

$$y = f(x) \tag{3}$$

- Parsimonious models
- Computation time for fit - determine model parameters from the data
- Choosing hyperparameters - free parameters

Example of Model Complexity and Overfitting



- Overfitting - low training error, high test error - lack of generalization

Regularization - Control the Complexity

- Fitting or training a model is an ill-posed problem
- Every model or algorithm requires a fit process to determine unknown parameters
- The fit process is eventually recast as an optimization problem
- Simpler model
- Sparse model
- Basic principle - cost function + a complexity penalty

Regularization - Early Stopping

- Regularization in time
- When the model performance doesn't improve on the validation set, stop
- Evaluate once more on test set to estimate generalization error
- Often used with neural networks and tree-based algorithms

Regularization - Control Model Complexity

$$y = f(x) \tag{4}$$

$$C(x, y) = \min_{f} \sum_{i=1}^{n} \|f(x_i) - y_i\|^2 + \lambda\|f\|^2 \tag{5}$$

- All parameters are driven to zero (but not all are non–zero)
- The underlying optimization problem can be solved with minimizers which require first and second derivatives
- These methods are sometimes faster than minizers which do not use gradient information
- There may be explicit matrix solutions

Regularization - Introducing Sparsity

- We want the model to have many hidden parameters zero
- Sometimes use the number of non–zero parameters
- Use the sum of the absolute value - $|\beta_1 + \beta_2 + \cdots + \beta_m|$
- This drives parameters to zero, except for a few
- Correlated features have a few representatives included, the others are not included

Regularization - ElasticNet

$$y = f(x) \tag{6}$$

$$C(x, y) = \min_f \sum_{i=1}^{n} \|f(x_i) - y_i\|^2 + \lambda \left(\|f\|^2 + (1 - \alpha) \sum |\beta_i| \right) \tag{7}$$

- Use the measures $\|.\|^2$ and $|.|$ together
- λ and α are hyperparameters that must be chosen via cross-validation or some other method
- Correlated features are assigned equal weights

The End - Thank You for Listening!