Model selection problems

Fraida Fund

Contents

Model selection problems

Bias variance tradeoff

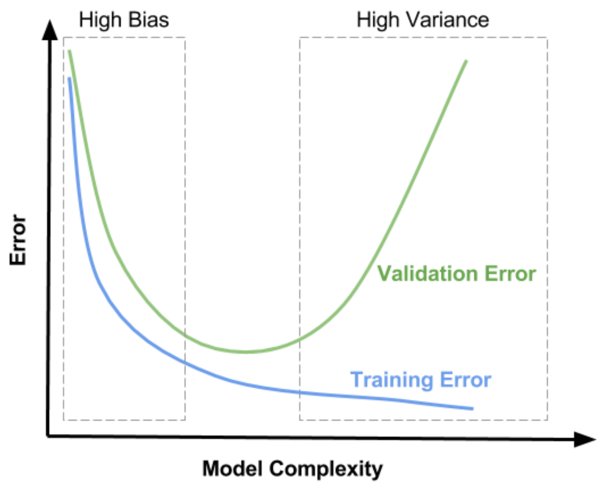

Figure 1: Bias variance tradeoff

Choosing model complexity

We need to select a model of appropriate complexity -

- what does that mean, and
- how do we select one?

Transformed linear models

Transformation of linear model

Standard linear model:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d$$

Transformed linear model:

$$\hat{y} = \beta_1 \phi_1(\mathbf{x}) + \cdots + \beta_p \phi_p(\mathbf{x})$$

Basis function

Each function

$$\phi_j(\mathbf{x}) = \phi_j(x_1, \cdots, x_d)$$

is called a **basis function**. These can be expressed in vector form:

$$\hat{y} = \boldsymbol{\phi}(\mathbf{x})\boldsymbol{\beta}$$

$$\boldsymbol{\phi}(\mathbf{x}) = [\phi_1(\mathbf{x}), \cdots, \phi_p(\mathbf{x})], \boldsymbol{\beta} = [\beta_1, \cdots, \beta_p]$$

Least squares for transformed linear models

Given data $(\mathbf{x_i}, y_i), i = 1, \cdots, N$:

$$A = \begin{bmatrix} \phi_1(\mathbf{x_1}) & \cdots & \phi_p(\mathbf{x_1}) \\ \vdots & \ddots & \vdots \\ \phi_1(\mathbf{x_N}) & \cdots & \phi_p(\mathbf{x_N}) \end{bmatrix}$$

Least squares fit is still $\widehat{\beta} = (A^T A)^{-1} A^T y$

Polynomial fitting

- Given data $(x_i, y_i), i = 1 \cdots, N$ (one feature)
- Polynomial model: $\hat{y} = \beta_0 + \beta_1 x + \cdots + \beta_d x^d$
- d is degree of polynomial, called **model order**. Given d, can get regression coefficients via LS

Transformed model for logistic regression

As with linear regression, can apply logistic regression to transformed features:

- $\phi(\mathbf{x}) = [\phi_1(\mathbf{x}), \ldots, \phi_p(\mathbf{x})]^T$
- Linear weights: $z_k = \sum_{j=1}^{p} W_{kj} \phi_j(\mathbf{x})$
- Softmax: $P(y = k|\mathbf{z}) = g_k(\mathbf{z}) = \frac{e^{z_k}}{\sum_\ell e^{z_\ell}}$

Logistic regression: illustration

Example: using non-linear features to classify data that is not linearly separable:

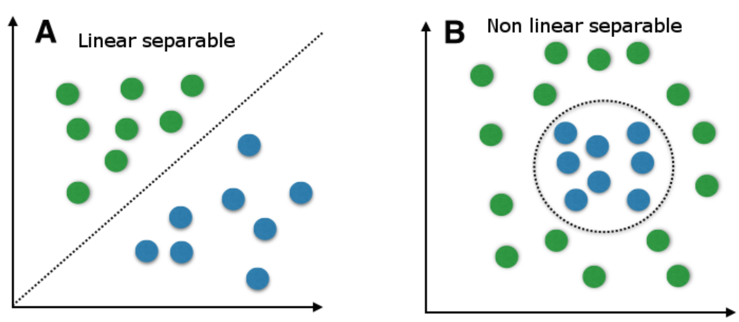

Figure 2: Non-linear data.

Logistic regression: example

$$\phi(\mathbf{x}) = [1, x_1, x_2, x_1^2, x_2^2]^T$$

Then can use $z = [-r^2, 0, 0, 1, 1]\phi(\mathbf{x}) = x_1^2 + x_2^2 - r^2$

Model order selection problem

Polynomial model: $\hat{y} = \beta_0 + \beta_1 x + \cdots + \beta_d x^d$

How can we select d when "true" model order is not known?

Model order illustrations

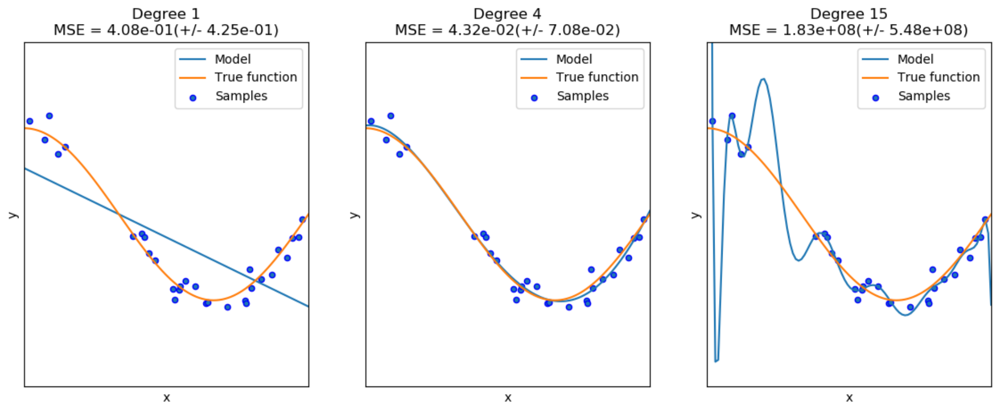

Figure 3: Model order selection; overfitting vs. underfitting

Using loss function for model order selection?

Suppose we would "search" over each possible d:

- Fit model of order d on draining data, get $\hat{\boldsymbol{\beta}}$
- Compute predictions on training data: $\hat{y}_i = \hat{\boldsymbol{\beta}}^T \mathbf{x_i}$
- Compute loss function (e.g. RSS) on training data: $RSS = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$
- Select d that minimizes loss
- Problem: loss function always decreasing with d (training error decreases with model complexity!)

Feature selection problem

Feature selection problem

- Linear model: $\hat{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d$
- Model target y as a function of features $\mathbf{x} = (x_1, \cdots, x_d)$
- Many features, only some are relevant
- High risk of overfitting if you use all features!
- Problem: fit a model with a small number of features

Feature selection problem - formal

Problem: given high dimensional data $\mathbf{X} \in R^{N \times p}$ and target variable y,

Select a subset of $k << p$ features, $\mathbf{X}_S \in R^{N \times k}$ that is most relevant to target y.

Motivation for feature selection problem

- Limited data
- Very large number of features
- Decrease variance

Limit on features for linear regression LS solution

For linear regression:

- We will have a unique solution to the least squares problem only if $A^T A$ is invertible.
- Solution is unique if $N \geq p$.

The unique solution exists only if the number of data samples for training (N) is greater than or equal to the number of parameters p.

Important applications for feature selection problem

- Document classification using "bag of words" - enumerate all words, represent each document using word count
- EEG - measure brain activity with electrodes, typically >10,000 "voxels" but only 100s of observations
- DNA MicroArray data - measures "expression" levels of large number of genes (~1000) but only a small number of data points (~100)

Decreasing variance for linear regression

For linear regression, when $N \geq p$,

$$Var = \frac{p}{N}\sigma_\epsilon^2$$

Variance increases linearly with number of parameters, inversely with number of samples.

(not derived in class, but read notes at home.)