6.867 Machine learning: lecture 4
Tommi S. Jaakkola
MIT CSAIL
tommi@csail.mit.edu

Models and accuracy
• We are interested in studying how the choice of inputs \( \{x_1, \ldots, x_n\} \) or, equivalently, \( X \), affects the accuracy of our regression model.

Polynomial regression
• Consider again a simple \( m^{th} \) degree polynomial regression model
  \[
y = w_0 + w_1 x + \ldots + w_m x^m + \epsilon, \quad \epsilon \sim N(0, \sigma^2)
\]
  where \( \sigma^2 \) is assumed fixed (known).

Topics
• Parameter uncertainty
  – regression model, underlying model
  – mean and variance of the ML estimator
• Active learning
  – measures of uncertainty
  – selection criteria, algorithms

Models and accuracy
• We are interested in studying how the choice of inputs \( \{x_1, \ldots, x_n\} \) or, equivalently, \( X \), affects the accuracy of our regression model.

Polynomial regression
• Consider again a simple \( m^{th} \) degree polynomial regression model
  \[
y = w_0 + w_1 x + \ldots + w_m x^m + \epsilon, \quad \epsilon \sim N(0, \sigma^2)
\]
  where \( \sigma^2 \) is assumed fixed (known).
• In this model the outputs \( \{y_1, \ldots, y_n\} \) corresponding to any inputs \( \{x_1, \ldots, x_n\} \) are generated according to
  \[
y = Xw + \epsilon, \quad \text{where}
  \[
y = \begin{bmatrix} y_1 \\
                        \vdots \\
                        y_n \end{bmatrix},
  X = \begin{bmatrix} 1 & x_1 & \cdots & x_1^m \\
                        \vdots & \vdots & \cdots & \vdots \\
                        1 & x_n & \cdots & x_n^m \end{bmatrix},
  \epsilon = \begin{bmatrix} \epsilon_1 \\
                        \vdots \\
                        \epsilon_n \end{bmatrix}
\]
  and \( \epsilon_i \sim N(0, \sigma^2) \), \( i = 1, \ldots, n \).
Models and accuracy
• We are interested in studying how the choice of inputs \( \{x_1, \ldots, x_n\} \) or, equivalently, \( X \), affects the accuracy of our regression model
• Our model for the outputs \( y = \{y_1, \ldots, y_n\} \) given \( X \) is
\[
y = Xw + e, \quad e \sim N(0, \sigma^2 I)
\]
• We assume also that the training outputs are actually generated by a model in this class with some fixed but unknown parameters \( w^* \) (same \( \sigma^2 \)):
\[
y = Xw^* + e, \quad e \sim N(0, \sigma^2 I)
\]
• We can now ask, for a given \( X \), how accurately we are able to recover the "true" parameters \( w^* \)

ML estimator, uncertainty
• The ML estimator \( \hat{w} \), viewed here as a function of the outputs \( y \) for a fixed \( X \), is given by
\[
\hat{w} = (X^T X)^{-1}X^T y
\]
• We need to understand how \( \hat{w} \) varies in relation to \( w^* \) when the outputs are generated according to
\[
y = Xw^* + e, \quad e \sim N(0, \sigma^2 I)
\]
• In the absence of noise \( e \), the ML estimator would recover \( w^* \) exactly (with only minor constraints on \( X \)):
\[
\hat{w} = (X^T X)^{-1}X^T (Xw^*) = (X^T X)^{-1}(X^T X)w^* = w^*
\]

ML estimator and noise
• In the presence of noise we can still use the fact that \( y = Xw^* + e \) to simplify the parameter estimates
\[
\hat{w} = (X^T X)^{-1}X^T y = (X^T X)^{-1}X^T (Xw^* + e) = (X^T X)^{-1}(X^T X)w^* + (X^T X)^{-1}X^T e = w^* + (X^T X)^{-1}X^T e
\]
So the ML estimate is the correct parameter vector plus an estimate based purely on noise.
**ML estimator**

- Since the ML estimator
  \[ \hat{w} = w^* + (X^T X)^{-1}X^T e \]
  is a linear function of normally distributed noise \( e \), it is also normally distributed.
- To fully characterize its distribution, given \( X \), we only need to evaluate its mean
  \[ \mu_{\hat{w}} = E\{ \hat{w} | X \} \]
  and covariance
  \[ C_{\hat{w}, \hat{w}} = E\{ (\hat{w} - \mu_{\hat{w}})(\hat{w} - \mu_{\hat{w}})^T | X \} \]

**ML estimator: covariance**

- We will again use the decomposition
  \[ \hat{w} = w^* + (X^T X)^{-1}X^T e \]
  and the fact that the mean is \( w^* \), and get
  \[
  E\{ (\hat{w} - w^*)(\hat{w} - w^*)^T | X \} = E\{ [(X^T X)^{-1}X^T e](X^T X)^{-1}X^T e]^T | X \} = E\{ [e^T X (X^T X)^{-1}]^T | X \} = (X^T X)^{-1}X^T e e^T X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}X^T X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}
  \]

**ML estimator: mean**

- Since the noise is zero mean by assumption, our parameter estimator is unbiased:
  \[
  E\{ \hat{w} | X \} = w^* + E\{ (X^T X)^{-1}X^T e | X \} = w^* + (X^T X)^{-1}X^T E\{ e | X \} = w^* + (X^T X)^{-1}X^T 0 = w^*
  \]

**ML estimator: summary**

- When the assumptions in the polynomial regression model are correct, the ML (least squares) estimator \( \hat{w} \), given \( X \), follows a simple Gaussian distribution:
  \[ \hat{w} \sim N( w^*, \sigma^2 (X^T X)^{-1} ) \]

  (the result naturally extends to any additive model)
- We can now study how the uncertainty (covariance) of this estimator depends on the choice of input points or \( X \)

**Topics**

- Parameter uncertainty
  - regression model, underlying model
  - mean and variance of the ML estimator
- Active learning
  - measures of uncertainty
  - selection criteria, algorithms

**Active learning**

- The ability to guide the selection of training inputs can substantially improve the accuracy of predictions when the data is otherwise limited
  - e.g., select specific documents to classify, faces to label, cars to test for fuel efficiency, etc.
- In active learning we try to optimize the selection of training inputs so as to maximally reduce model/prediction uncertainty
Active regression
- For any set of training inputs \( X \) the resulting uncertainty about the parameters is characterized by the covariance matrix \( \sigma^2(X^TX)^{-1} \) of the Gaussian distribution
  \[
  \hat{w} \sim N(w^*, \sigma^2(X^TX)^{-1})
  \]
  Note that the covariance matrix does not depend on the training outputs!
- We’d like to select input points, specify \( X \), so as to minimize any residual “uncertainty”; need to define exactly how to measure uncertainty based on the covariance.

Determinant as a measure of “volume”
- Any covariance matrix has an eigen-decomposition:
  \[
  C = R \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_m^2 \end{bmatrix} R^T
  \]
  where the orthonormal rotation matrix \( R \) specifies the principal axes of variation and each eigenvalue \( \sigma_i^2 \) gives the variance along one of the principal directions.
- The “volume” of a Gaussian distribution is a function of only \( \sigma_i^2 \), \( i = 1, \ldots, m \). Specifically
  \[
  \text{“volume”} \propto \prod_{i=1}^m \sigma_i = \sqrt{\det C}
  \]

Parameter uncertainty
- Determinant of the covariance matrix is one possible measure of uncertainty, capturing the “volume” of variation around the mean.
- We can therefore find \( n \) inputs \( x_1, \ldots, x_n \), which determine the matrix \( X \), so as to minimize the determinant of the covariance matrix (\( \sigma^2 \) only affects the overall scaling, not the choice of points):
  \[
  \det \left[ (X^TX)^{-1} \right]
  \]
- Note that since the covariance does not depend on the training outputs, we can select the inputs either sequentially or prior to seeing any outputs.

Determinant criterion: example
- 1st order polynomial regression within \( x \in [-1, 1] \)
  \[
  f(x; w) = w_0 + w_1x
  \]
  What are the first two points that would we select?

Determinant criterion: example
- 2nd order polynomial regression within \( x \in [-1, 1] \)
  \[
  f(x; w) = w_0 + w_1x + w_2x^2
  \]
  What the first three points that we would select?
**Determinant criterion: example**

- 2nd order polynomial regression within $x \in [-1, 1]$
  
  \[ f(x; w) = w_0 + w_1 x + w_2 x^2 \]
  
  What is the first three points that we would select?

![Graph showing a polynomial function](image)

\[ x_1 = -1, x_2 = 0, x_3 = 1 \]

---

**Sequential selection**

- The determinant criterion is based on the uncertainty in the parameter values, not directly that of the predictions
- We can devise a sequential selection criterion that aims to minimize the variance of the predictions directly
- For example: the prediction at a new point $x$ is
  
  \[ f(x; \hat{w}) = \hat{w}_0 + \hat{w}_1 x = \begin{bmatrix} 1 \\ x \end{bmatrix}^T \hat{w}, \]
  
  with variance
  
  \[
  Var \{ f(x; \hat{w}) \} = \begin{bmatrix} 1 \\ x \end{bmatrix}^T C_{\hat{w}, \hat{w}} \begin{bmatrix} 1 \\ x \end{bmatrix}
  \]
  
  \[ = \sigma^2 \begin{bmatrix} 1 \\ x \end{bmatrix}^T (X^T X)^{-1} \begin{bmatrix} 1 \\ x \end{bmatrix} \]

---

**Sequential selection cont’d**

\[
Var \{ f(x; \hat{w}) \} = \sigma^2 \begin{bmatrix} 1 \\ x \end{bmatrix}^T (X^T X)^{-1} \begin{bmatrix} 1 \\ x \end{bmatrix}
\]

- $\sigma^2$ only affects the overall scale (set to 1 from hereafter)
- the variance is a function of both the query point $x$ and the past inputs or $X$

**Sequential selection: example**

- 2nd order polynomial regression within $x \in [-1, 1]$
  
  \[ f(x; \hat{w}) = \hat{w}_0 + \hat{w}_1 x + \hat{w}_2 x^2 \]
  
  A priori selected inputs $x_1 = -1, x_2 = 0, x_3 = 1$.

\[
Var \{ f(x; \hat{w}) \} = \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix}^T (X^T X)^{-1} \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix}
\]

where \[ X = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \end{bmatrix} \]

---

**Example cont’d**

![Graph showing a polynomial function](image)

---

**Sequential selection: properties**

- In the linear/additive regression context the prediction variance is uniformly non-increasing
  
  \[ C = (X^T X)^{-1} \quad \text{covariance of } \hat{w} \]
  
  \[ A = (X^T X) \quad \text{inverse covariance} \]

\[
Var \{ f(x; \hat{w}) \} = \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix}^T C \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix} = \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix}^T A^{-1} \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix}
\]

It suffices to show that the eigenvalues of $A$ can only increase (or remain the same) as a result of adding new inputs.
Brief derivation
Suppose we add any valid input $x'$,

$$A' = \begin{bmatrix} 1 & x' & x'^2 \\ x & X \end{bmatrix}^T \begin{bmatrix} 1 & x' & x'^2 \\ x & X \end{bmatrix}$$

$$= X^T X + \begin{bmatrix} 1 & x' & x'^2 \\ x' & x'^2 \end{bmatrix}^T \begin{bmatrix} 1 & x' & x'^2 \\ x' & x'^2 \end{bmatrix}$$

$$= A + \begin{bmatrix} 1 \\ x' \\ x'^2 \end{bmatrix}^T \begin{bmatrix} 1 \\ x' \\ x'^2 \end{bmatrix}$$

In other words, we add to $A$ a matrix whose eigenvalues are all non-negative $\Rightarrow$ eigenvalues of $A$ are non-decreasing