6.867 Machine learning: lecture 4

Tommi S. Jaakkola
MIT CSAIL
tommi@csail.mit.edu

Topics

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

Polynomial regression

- Consider again a simple m^{th} degree polynomial regression model

$$y = w_0 + w_1 x + \ldots + w_m x^m + \epsilon, \;\; \epsilon \sim N(0, \sigma^2)$$

where σ^2 is assumed fixed (known).

Polynomial regression

- Consider again a simple m^{th} degree polynomial regression model

$$y = w_0 + w_1 x + \ldots + w_m x^m + \epsilon, \;\; \epsilon \sim N(0, \sigma^2)$$

where σ^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

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{e}, \quad \text{where}$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \cdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 & \ldots & x_1^m \\ \cdots & \cdots & \cdots & \\ 1 & x_n & \ldots & x_n^m \end{bmatrix}, \quad \mathbf{e} = \begin{bmatrix} \epsilon_1 \\ \cdots \\ \epsilon_n \end{bmatrix}$$

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, $\mathbf{X}$, affects the accuracy of our regression model

Models and accuracy

- We are interested in studying how the choice of inputs $\{x_1, \ldots, x_n\}$ or, equivalently, $\mathbf{X}$, affects the accuracy of our regression model
- Our model for the outputs $\mathbf{y} = \{y_1, \ldots, y_n\}$ given $\mathbf{X}$ is

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{e}, \quad \mathbf{e} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

Models and accuracy

- We are interested in studying how the choice of inputs $\{x_1, \ldots, x_n\}$ or, equivalently, $\mathbf{X}$, affects the accuracy of our regression model
- Our model for the outputs $\mathbf{y} = \{y_1, \ldots, y_n\}$ given $\mathbf{X}$ is

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{e}, \quad \mathbf{e} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

- We assume also that the training outputs are actually generated by a model in this class with some fixed but unknown parameters $\mathbf{w}^*$ (same σ^2):

$$\mathbf{y} = \mathbf{X}\mathbf{w}^* + \mathbf{e}, \quad \mathbf{e} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

Models and accuracy

- We are interested in studying how the choice of inputs $\{x_1, \ldots, x_n\}$ or, equivalently, $\mathbf{X}$, affects the accuracy of our regression model
- Our model for the outputs $\mathbf{y} = \{y_1, \ldots, y_n\}$ given $\mathbf{X}$ is

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{e}, \quad \mathbf{e} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

- We assume also that the training outputs are actually generated by a model in this class with some fixed but unknown parameters $\mathbf{w}^*$ (same σ^2):

$$\mathbf{y} = \mathbf{X}\mathbf{w}^* + \mathbf{e}, \quad \mathbf{e} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

- We can now ask, for a given $\mathbf{X}$, how accurately we are able to recover the "true" parameters $\mathbf{w}^*$

ML estimator, uncertainty

- The ML estimator $\hat{\mathbf{w}}$, viewed here as a function of the outputs $\mathbf{y}$ for a fixed $\mathbf{X}$, is given by

$$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

ML estimator, uncertainty

- The ML estimator $\hat{\mathbf{w}}$, viewed here as a function of the outputs $\mathbf{y}$ for a fixed $\mathbf{X}$, is given by

$$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

- We need to understand how $\hat{\mathbf{w}}$ varies in relation to $\mathbf{w}^*$ when the outputs are generated according to

$$\mathbf{y} = \mathbf{X}\mathbf{w}^* + \mathbf{e}, \quad \mathbf{e} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

ML estimator, uncertainty

- The ML estimator $\hat{\mathbf{w}}$, viewed here as a function of the outputs $\mathbf{y}$ for a fixed $\mathbf{X}$, is given by

$$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

- We need to understand how $\hat{\mathbf{w}}$ varies in relation to $\mathbf{w}^*$ when the outputs are generated according to

$$\mathbf{y} = \mathbf{X}\mathbf{w}^* + \mathbf{e}, \quad \mathbf{e} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

- In the absence of noise $\mathbf{e}$, the ML estimator would recover $\mathbf{w}^*$ exactly (with only minor constraints on $\mathbf{X}$):

$$\begin{aligned} \hat{\mathbf{w}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\mathbf{w}^*) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})\mathbf{w}^* \\ &= \mathbf{w}^* \end{aligned}$$

ML estimator and noise

- In the presence of noise we can still use the fact that $\mathbf{y} = \mathbf{X}\mathbf{w}^* + \mathbf{e}$ to simplify the parameter estimates

$$\begin{aligned} \hat{\mathbf{w}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\mathbf{w}^* + \mathbf{e}) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})\mathbf{w}^* + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{e} \\ &= \mathbf{w}^* + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{e} \end{aligned}$$

So the ML estimate is the correct parameter vector plus an estimate based purely on noise.

ML estimator

- Since the ML estimator

$$\hat{\mathbf{w}} \;=\; \mathbf{w}^* + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{e}$$

  is a linear function of normally distributed noise $\mathbf{e}$, it is also normally distributed.

- To fully characterize its distribution, given $\mathbf{X}$, we only need to evaluate its

  mean

$$\mu_{\hat{\mathbf{w}}} = E\{\, \hat{\mathbf{w}} \,|\mathbf{X}\}$$

  and covariance

$$C_{\hat{\mathbf{w}},\hat{\mathbf{w}}} = E\{\, (\hat{\mathbf{w}} - \mu_{\hat{\mathbf{w}}})(\hat{\mathbf{w}} - \mu_{\hat{\mathbf{w}}})^T \,|\mathbf{X}\}$$

ML estimator: mean

- Since the noise is zero mean by assumption, our parameter estimator is *unbiased*:

$$\begin{aligned} E\{\, \hat{\mathbf{w}} \,|\mathbf{X}\} &= \mathbf{w}^* + E\left\{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{e}\,|\mathbf{X}\right\} \\ &= \mathbf{w}^* + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E\left\{\mathbf{e}\,|\mathbf{X}\right\} \\ &= \mathbf{w}^* + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{0} \\ &= \mathbf{w}^* \end{aligned}$$

ML estimator: covariance

- We will again use the decomposition

$$\hat{\mathbf{w}} = \mathbf{w}^* + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{e}$$

  and the fact that the mean is $\mathbf{w}^*$, and get

$$\begin{aligned} & E\left\{\, (\hat{\mathbf{w}} - \mathbf{w}^*)(\hat{\mathbf{w}} - \mathbf{w}^*)^T \,|\mathbf{X}\right\} \\ &= E\left\{\, \left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{e}\right]\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{e}\right]^T \,|\mathbf{X}\right\} \\ &= E\left\{\, \left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{e}\right]\left[\mathbf{e}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\right] |\mathbf{X}\right\} \\ &= \left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right] E\left\{\,\mathbf{e}\mathbf{e}^T\,|\mathbf{X}\right\} \left[\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\right] \\ &= \left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right] \sigma^2\mathbf{I} \left[\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\right] \\ &= \sigma^2 \left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\right] \\ &= \sigma^2\,(\mathbf{X}^T\mathbf{X})^{-1} \end{aligned}$$

ML estimator: summary

- When the assumptions in the polynomial regression model are correct, the ML (least squares) estimator $\hat{\mathbf{w}}$, given $\mathbf{X}$, follows a simple Gaussian distribution:

$$\hat{\mathbf{w}} \sim N(\,\mathbf{w}^*,\, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\,)$$

  (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 $\mathbf{X}$

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 $\mathbf{X}$ the resulting uncertainty about the parameters is characterized by the covariance matrix $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$ of the Gaussian distribution

$$\hat{\mathbf{w}} \sim N(\, \mathbf{w}^*, \, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \,)$$

  Note that the covariance matrix does not depend on the training outputs!

- We'd like to select input points, specify $\mathbf{X}$, so as to minimize any residual "uncertainty"; need to define exactly how to measure uncertainty based on the covariance

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 $\mathbf{X}$, so as to minimize the determinant of the covariance matrix (σ^2 only affects the overall scaling, not the choice of points):

$$\text{det}\left[\, (\mathbf{X}^T\mathbf{X})^{-1} \,\right]$$

- 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 as a measure of "volume"

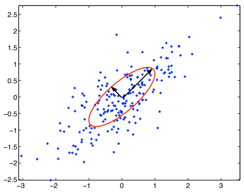
- Any covariance matrix has an eigen-decomposition:

$$\mathbf{C} = \mathbf{R} \begin{bmatrix} \sigma_1^2 & & \\ & \ldots & \\ & & \sigma_m^2 \end{bmatrix} \mathbf{R}^T$$

  where the orthonormal rotation matrix $\mathbf{R}$ specifies the principal axes of variation and each eigenvalue σ_i^2 gives the variance along one of the principal directions

- The "volume" of a Gaussian distribution is a function of only σ_i^2, $i = 1, \ldots, m$. Specifically

$$\text{"volume"} \propto \prod_{i=1}^{m} \sigma_i = \sqrt{\det C}$$



Determinant criterion: example

- 1st order polynomial regression within $x \in [-1, 1]$

$$f(x; \mathbf{w}) = w_0 + w_1 x$$

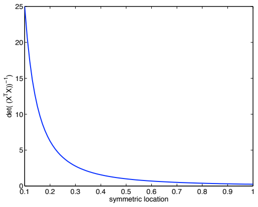
- What are the first two points that would we select?

Determinant criterion: example

- 1st order polynomial regression within $x \in [-1, 1]$

$$f(x; \mathbf{w}) = w_0 + w_1 x$$

- What are the first two points that would we select?



$$x_1 = 1, x_2 = -1$$

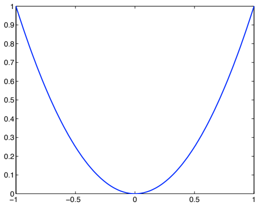
(the two points have to be symmetric around zero)

Determinant criterion: example

- 2nd order polynomial regression within $x \in [-1, 1]$

$$f(x; \mathbf{w}) = w_0 + w_1 x + w_2 x^2$$

What the first three points that we would select?

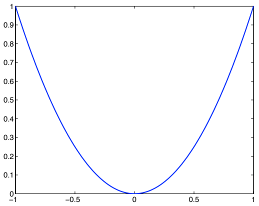


Determinant criterion: example

- 2nd order polynomial regression within $x \in [-1, 1]$

$$f(x; \mathbf{w}) = w_0 + w_1 x + w_2 x^2$$

What the first three points that we would select?



$$x_1 = -1, x_2 = 0, x_3 = 1$$

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

$$f(x; \hat{\mathbf{w}}) \; = \; \hat{w}_0 + \hat{w}_1 x = \begin{bmatrix} 1 \\ x \end{bmatrix}^T \hat{\mathbf{w}},$$

with variance

$$\begin{aligned} Var\{\, f(x; \hat{\mathbf{w}})\,\} \; &= \; \begin{bmatrix} 1 \\ x \end{bmatrix}^T C_{\hat{\mathbf{w}},\hat{\mathbf{w}}} \begin{bmatrix} 1 \\ x \end{bmatrix} \\ &= \; \sigma^2 \begin{bmatrix} 1 \\ x \end{bmatrix}^T (\mathbf{X}^T\mathbf{X})^{-1} \begin{bmatrix} 1 \\ x \end{bmatrix} \end{aligned}$$

Sequential selection cont’d

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

- σ^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 $\mathbf{X}$

- Assuming the input points are contained within, e.g., an interval $\mathcal{X}$, we can select the next input to be the point of most uncertain prediction:

$$x^{new} = \operatorname*{argmax}_{x \in \mathcal{X}} \left\{ Var\{\, f(x; \hat{\mathbf{w}})\,\} \right\}$$

Sequential selection: example

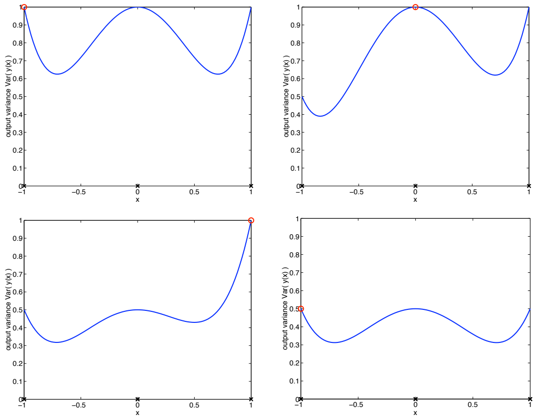
- 2nd order polynomial regression within $x \in [-1, 1]$

$$f(x; \hat{\mathbf{w}}) = \hat{w}_0 + \hat{w}_1 x + \hat{w}_2 x^2$$

A priori selected inputs $x_1 = -1, x_2 = 0, x_3 = 1$.

$$\begin{aligned} Var\{\, f(x; \hat{\mathbf{w}})\,\} \; &= \; \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix}^T (\mathbf{X}^T\mathbf{X})^{-1} \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix} \\ \text{where } \mathbf{X} \; &= \; \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \cdots & \cdots & \cdots \end{bmatrix} \end{aligned}$$

Example cont’d



Sequential selection: properties

- In the linear/additive regression context the prediction variance is uniformly non-increasing

$$\begin{aligned} \mathbf{C} \; &= \; (\mathbf{X}^T\mathbf{X})^{-1} \quad \text{covariance of } \hat{\mathbf{w}} \\ \mathbf{A} \; &= \; (\mathbf{X}^T\mathbf{X}) \quad \text{inverse covariance} \end{aligned}$$

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

It suffices to show that the eigenvalues of $\mathbf{A}$ can only increase (or remain the same) as a result of adding new inputs.

Brief derivation

Suppose we add any valid input x',

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

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