Machine learning: lecture 4

Tommi S. Jaakkola
MIT AI Lab
tommi@ai.mit.edu

Topics

- Active learning and regression
  - formulation
  - selection criteria
  - examples

Active learning: rules of the game

- Supervised learning:
  - (input,output) pairs are sampled from an *unknown joint* distribution $P(x, y)$
- Active supervised learning:
  - We select the input examples and the corresponding outputs are sampled from an *unknown conditional* distribution $P(y|x)$

Active learning

- Why active learning?
  - we often need dramatically fewer training examples; the time/cost of getting enough training examples may be otherwise prohibitive
- Dangers of (this type of) active learning
  - since we select the inputs, we may focus on inputs that are unimportant, rare, or even invalid

Active learning

- We need to decide:
  1. the function class (the result will be highly dependent on what we wish to learn)
  2. the selection criterion, i.e., how we decide which inputs are worth querying
  3. how to apply the selection criterion (sequential or batch)
- Function class: we'll focus on linear/polynomial regression

$$y = w_0 + w_1 x + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

Active linear regression

- We perform the selection of inputs to uncover the (assumed) "true" underlying linear relation:

$$\begin{bmatrix} y_1 \\ \cdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \cdots & \cdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} w_0^* \\ w_1^* \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \cdots \\ \epsilon_n \end{bmatrix}$$

$$\mathbf{y} = \mathbf{X}\mathbf{w}^* + \epsilon$$

where $\epsilon_i \sim N(0, \sigma^2)$.

- We need to first understand how our parameter estimates relate to $\mathbf{w}^*$ as a function of inputs

Properties of regression models

- The outputs corresponding to the inputs arranged in $\mathbf{X}$ are assumed to be generated according to:

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

- The resulting parameter estimates, $\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, based on the same inputs $\mathbf{X}$ and sampled outputs $\mathbf{y}$ are normally distributed:

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

Active learning: selection criterion

- Two main types of selection criteria
  1. select inputs so as to minimize some measure of uncertainty in the *parameters*
  2. select inputs to minimize the uncertainty in the *predicted outputs*

- Two main ways of applying such criteria
  1. *batch* – all the inputs are chosen prior to seeing any responses
  2. *sequential* – the next query input is chosen with the full knowledge of all the responses so far

Batch selection, parameter criterion

We have to select the input examples prior to seeing any outputs

- We wish to find n inputs $x_1, \ldots, x_n$ (which determine the matrix $\mathbf{X}$) so as to minimize a measure of uncertainty in the resulting parameters $\hat{\mathbf{w}}$

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

- For example, we can find the inputs that minimize the determinant of the covariance matrix

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

Determinant as a measure of "volume"

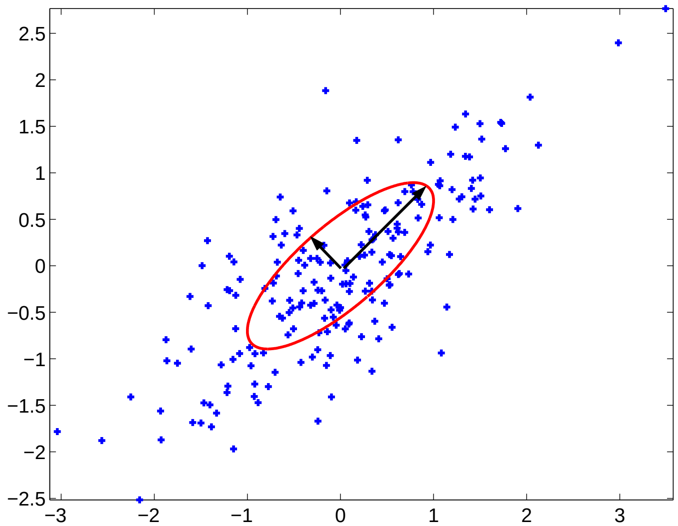
- Any covariance matrix has an eigen-decomposition:

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

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

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

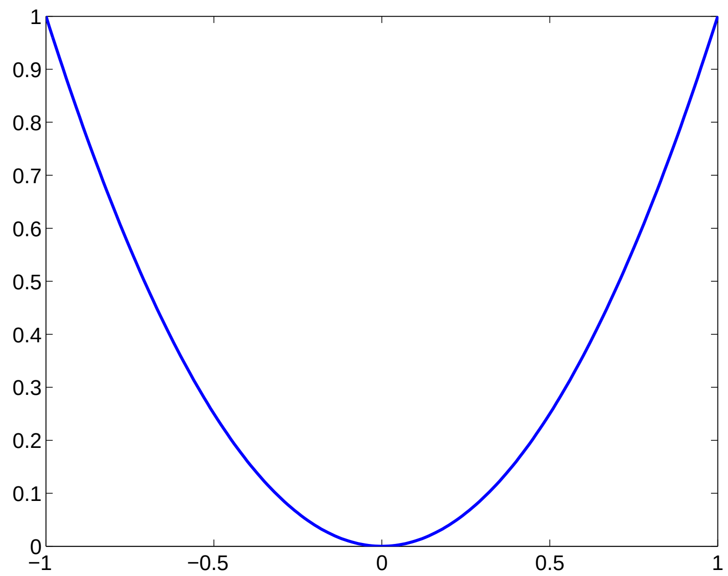


Determinant criterion: example

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

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

For $n = 4$, what points would we select?

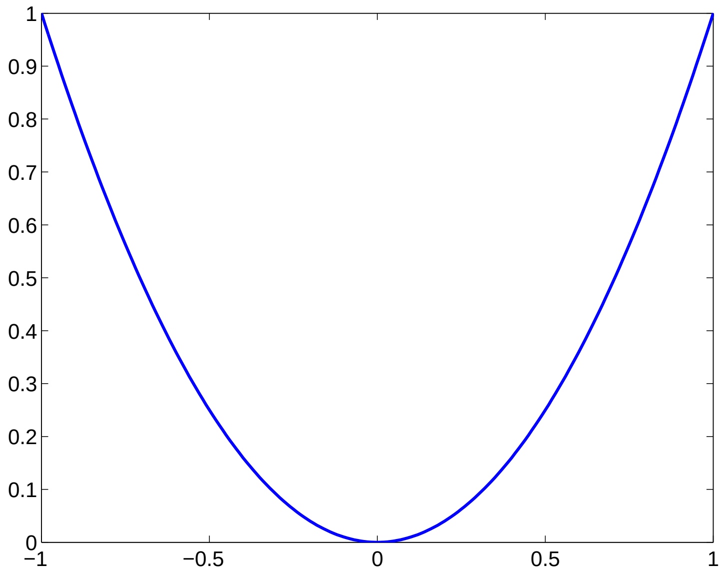


Determinant criterion: example

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

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

For $n = 4$, what points would we select?



$$x_1 = -1, x_2 = 0, x_3 = 0, x_4 = 1$$

Sequential selection, uncertainty in predictions

The next input is chosen on the basis of all the information available so far

- The prediction at a new point x is

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

The variance in this prediction (due to the noise in the outputs observed so far) is

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

Sequential selection cont'd

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

– the noise variance σ^2 only affects the overall scale (set to 1 from hereafter)
– the variance is a function of previously chosen inputs, not outputs!

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

$$x^{new} = \arg\max_{x\in\mathcal{X}} \left\{ Var\{\,\hat{y}(x)\,\} \right\}$$

Sequential selection: example

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

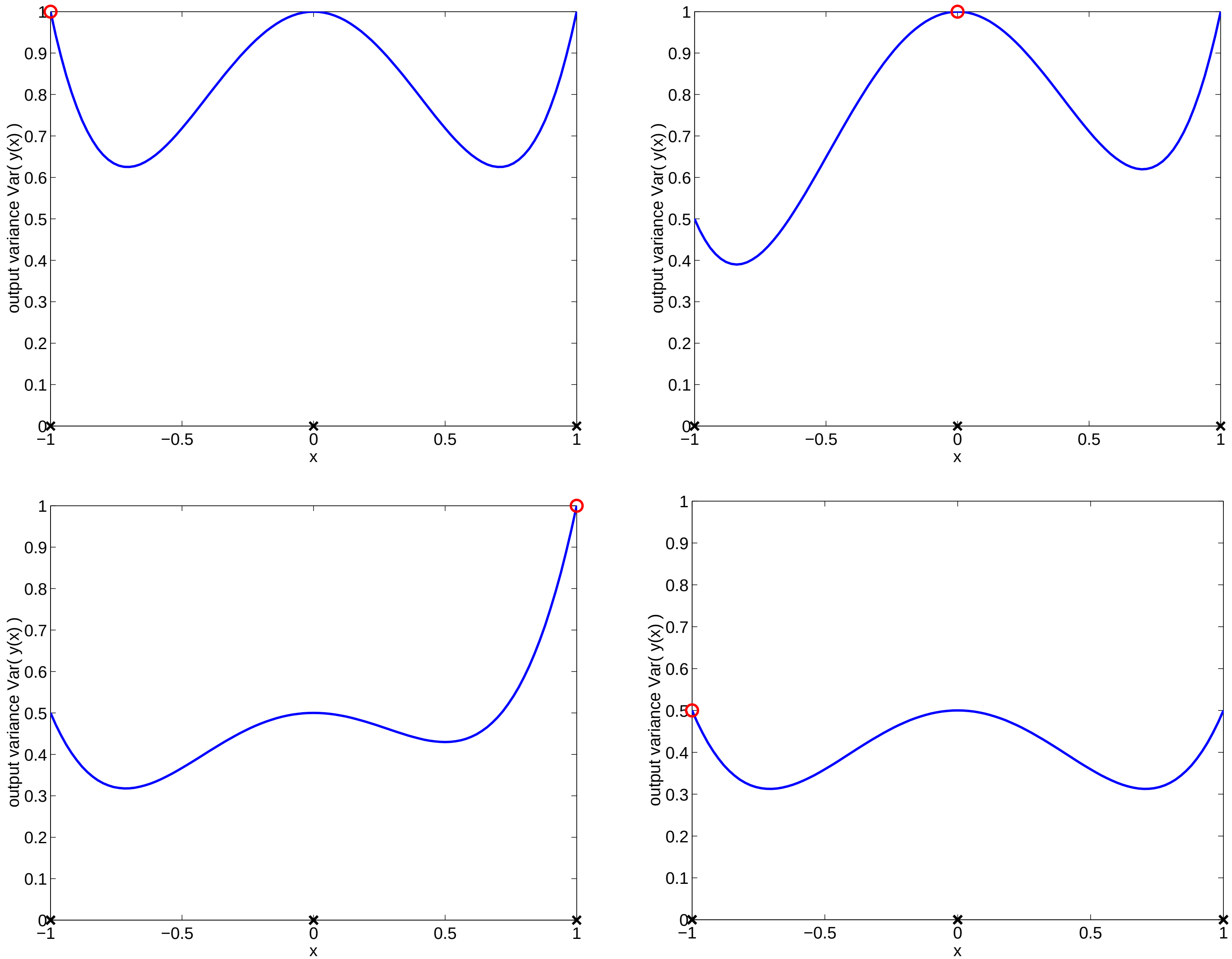
$$\hat{y}(x) = \hat{w}_0 + \hat{w}_1 x + \hat{w}_2 x^2$$

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

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

$$\text{where } \mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \cdots & \cdots & \cdots \end{bmatrix}$$

Example cont'd



Sequential selection: properties

- In the linear/additive regression context the variance cannot increase anywhere as new points are added

$$\mathbf{C} = (\mathbf{X}^T\mathbf{X})^{-1} \quad \text{covariance of } \hat{\mathbf{w}}$$

$$\mathbf{A} = (\mathbf{X}^T\mathbf{X}) \quad \text{inverse covariance}$$

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

The variance never increases for any point x if the eigenvalues of the inverse covariance matrix $\mathbf{A}$ increase (or stay the same) as we add new points

Brief derivation

New query point 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 $\Rightarrow$ eigenvalues of $\mathbf{A}$ are non-decreasing

Active learning more generally

- To perform active learning we have to evaluate "the value of new information", i.e., how much we expect to gain from querying another response
- Such calculations can be done in the context of almost any learning task

we will revisit the issue later on in the course ...