Machine learning: lecture 4

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

Topics

- Properties of estimators
  - bias, variance
- Active learning and regression
  - motivation
  - selection criteria
  - examples

Review: statistical regression models

- We specify a probabilistic model for how outputs are generated from the inputs

$$\begin{aligned} \textbf{Output} &= \textbf{function} + \textbf{noise} \\ y &= f(\mathbf{x}; \mathbf{w}) + \epsilon, \quad \epsilon \sim N(0, \sigma^2) \end{aligned}$$

Example: simple linear regression

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

$$p(y|x, \mathbf{w}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(y - w_0 - w_1 x)^2\}$$

Review: ML estimation

- When the noise is assumed to be zero mean Gaussian, maximum likelihood setting of the linear regression parameters $\mathbf{w} = [w_0, w_1]^T$ reduces to least squares fitting:

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

where

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \cdots & \cdots \\ 1 & x_n \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \cdots \\ y_n \end{bmatrix}, \quad \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

Review: ML estimation

- When the noise is assumed to be zero mean Gaussian, maximum likelihood setting of the linear regression parameters $\mathbf{w} = [w_0, w_1]^T$ reduces to least squares fitting:

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

where

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \cdots & \cdots \\ 1 & x_n \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \cdots \\ y_n \end{bmatrix}, \quad \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

- We can study the estimator $\hat{\mathbf{w}}$ further if we assume that the linear model is indeed correct, i.e.,. the outputs were generated according to $y = w_0^* + w_1^* x + \epsilon$ with some unknown parameters $\mathbf{w}^* = [w_0^*, w_1^*]^T$.

Properties of the ML estimator

- Major assumption: $\mathbf{y} = \mathbf{X}\mathbf{w}^* + \mathbf{e}$, where $\mathbf{e} = [\epsilon_1, \ldots, \epsilon_n]^T$, $\epsilon_i \sim N(0, \sigma^2)$.
- We keep the training inputs or $\mathbf{X}$ fixed and study how $\hat{\mathbf{w}}$ varies if we resample the corresponding outputs

  Bias: whether $\hat{\mathbf{w}}$ deviates from $\mathbf{w}^*$ on average

  $$E\{\, \hat{\mathbf{w}} \,|\mathbf{X}\} - \mathbf{w}^*$$

  Variance (covariance): how much $\hat{\mathbf{w}}$ varies around its mean

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

  where $\mu = E\{\, \hat{\mathbf{w}} \,|\mathbf{X}\}$.

ML estimator

- We can now use the fact that $\mathbf{y} = \mathbf{X}\mathbf{w}^* + \mathbf{e}$ to simplify the parameter estimates

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

- The parameter estimate based on the sampled data is therefore the correct parameter plus estimate based purely on noise.

Estimator: bias

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

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

  where the conditional expectation is over the noisy outputs while keeping the inputs $x_1, \ldots, x_n$, or $\mathbf{X}$, fixed.

Estimator: covariance

- We can also evaluate the (conditional) covariance of the parameters, i.e., how the individual parameters co-vary due to the noise in the outputs:

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

ML estimator: summary

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

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

  (the result naturally extends to any additive model)

- The above Gaussian distribution summarizes the uncertainty that we have about the parameters based on a training set $D_n = \{(x_1, y_1) \ldots, (x_n, y_n)\}$

Active learning

- In many cases we can actually select the training inputs for which we will receive the corresponding outputs
  - e.g., experiment design, querying user feedback, etc.
- Why useful?

  For example, the uncertainty we have about the parameters depends on the training inputs; we should be able to select inputs to maximally reduce this uncertainty

Active learning: value of new information

- In order to appropriately select training inputs we need to be able to

  1. predict possible output values for any (valid) input
  2. gauge the effect of including the input and the (predicted) output in the training set

  and we need a *selection criterion* that measures the value of this new information.

- Some immediate dangers to think about:
  - our model (e.g., linear) may be incorrect
  - we may focus on inputs that are unimportant, rare, or even invalid

Active linear regression

- We assume as before that for any set of inputs $x_1, \ldots, x_n$ or $\mathbf{X}$ the outputs are generated according to

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

- The resulting parameter estimate $\hat{\mathbf{w}}$ is normally distributed according to

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

- We can use this information to
  1. select inputs so as to minimize some measure of uncertainty in the *parameters*
  2. select inputs to minimize the uncertainty in *predicted outputs*

Parameter criterion

We select the inputs prior to seeing any outputs (this is possible but not necessary)

- 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"

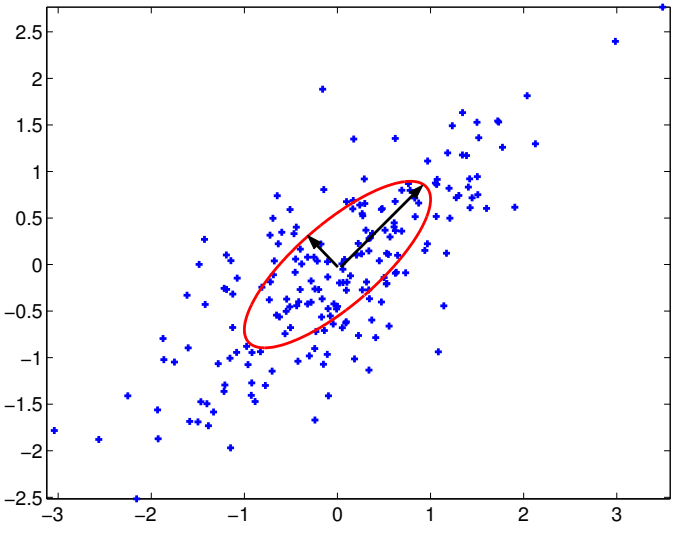
- Any covariance matrix has an eigen-decomposition:

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

  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, $i = 1, \dots, m$. Specifically

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

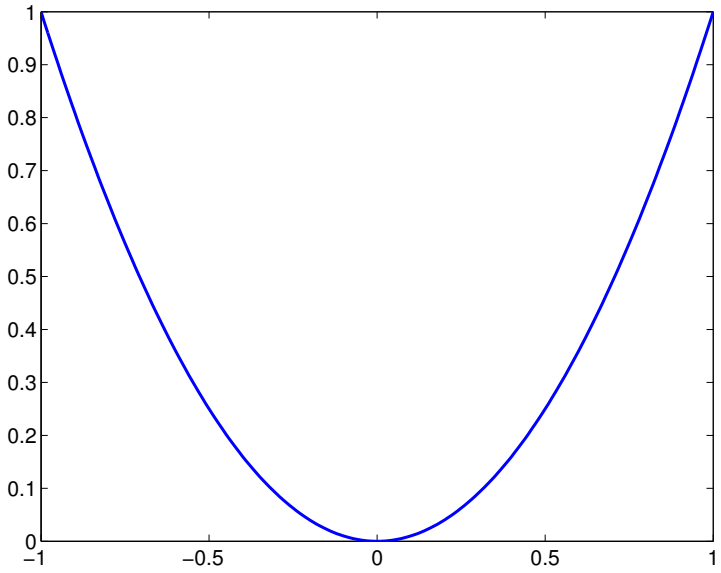


Determinant criterion: example

- 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 end up selecting?

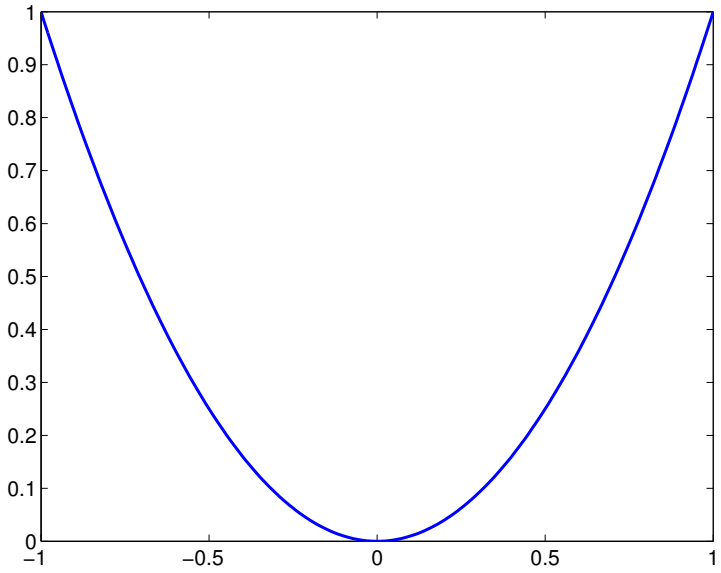


Determinant criterion: example

- 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 end up selecting?



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

Prediction criterion

We will choose the next input sequentially 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 $x \in [-1, 1]$

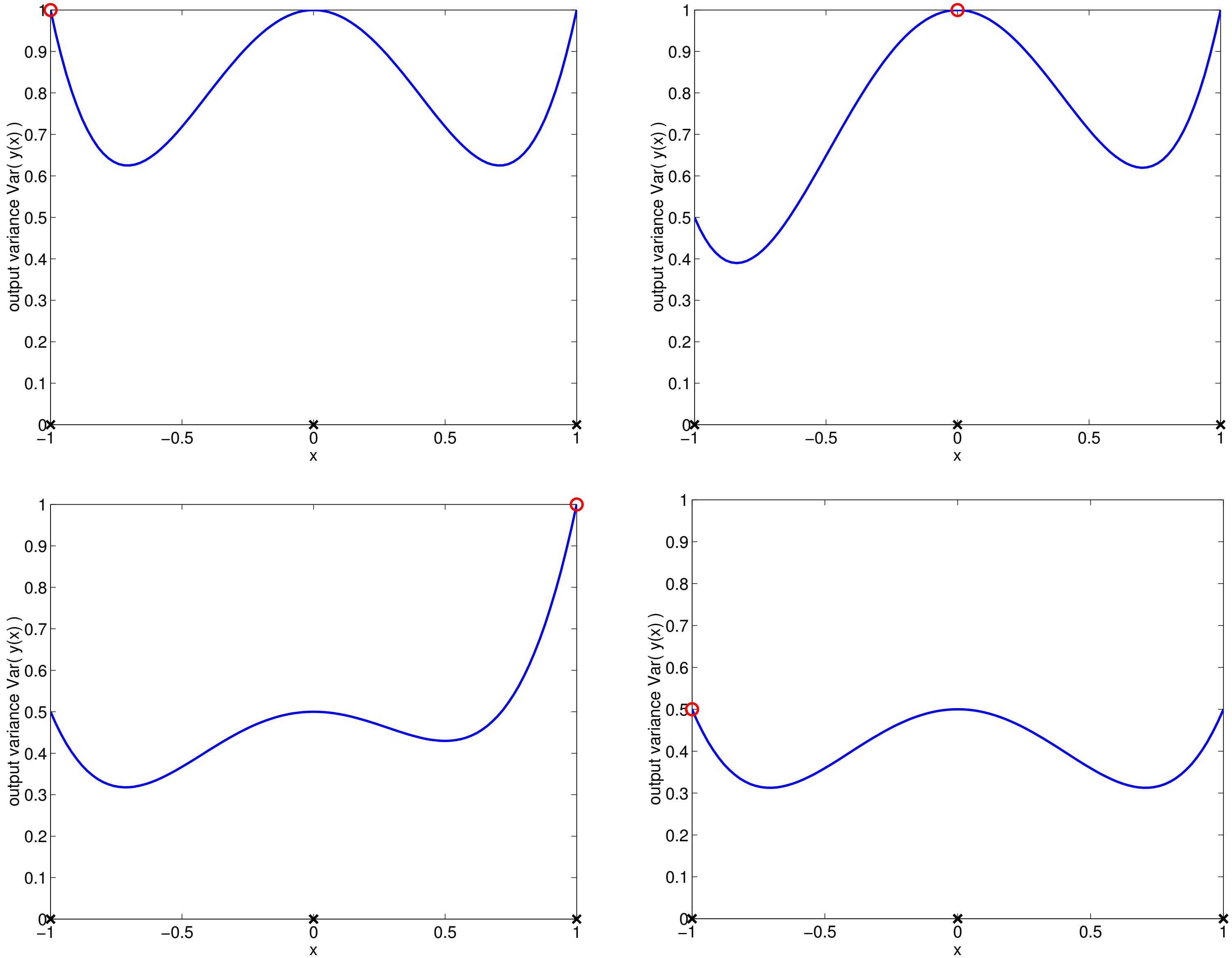
$$\hat{y}(x) = \hat{w}_0 + \hat{w}_1 x + \hat{w}_2 x^2$$

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

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

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

Example cont'd



Sequential selection: properties

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

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

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

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 ...