Machine learning: lecture 2

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

Topics

- Regression
  - examples, assumptions, abstraction
- Linear regression
  - estimation, properties
  - generalization concepts

Regression problems

- The goal is to make quantitative (real valued) predictions on the basis of a (vector of) features or attributes
- Examples: house prices, stock values, survival time, fuel efficiency of cars, etc.
- what can we assume about the problem? how do we formalize the regression problem? how do we evaluate predictions?

A generic regression problem

- The input attributes are given as fixed length vectors $\mathbf{x} = [x_1, \ldots, x_d]^T$, where each component such as x_i may be discrete or real valued.
- The outputs are assumed to be real valued $y \in \mathcal{R}$ (the values of actual outputs such as prices may be more restricted)

A generic regression problem

- The input attributes are given as fixed length vectors $\mathbf{x} = [x_1, \ldots, x_d]^T$, where each component such as x_i may be discrete or real valued.
- The outputs are assumed to be real valued $y \in \mathcal{R}$ (the values of actual outputs such as prices may be more restricted)
- We have access to a set of n *training examples*, $D_n = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, sampled independently at random from some fixed but unknown distribution $P(\mathbf{x}, y)$

A generic regression problem

- The input attributes are given as fixed length vectors $\mathbf{x} = [x_1, \ldots, x_d]^T$, where each component such as x_i may be discrete or real valued.
- The outputs are assumed to be real valued $y \in \mathcal{R}$ (the values of actual outputs such as prices may be more restricted)
- We have access to a set of n *training examples*, $D_n = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, sampled independently at random from some fixed but unknown distribution $P(\mathbf{x}, y)$
- The goal is to minimize the prediction error/loss on new examples $(\mathbf{x}, y)$ drawn at random from the same $P(\mathbf{x}, y)$.

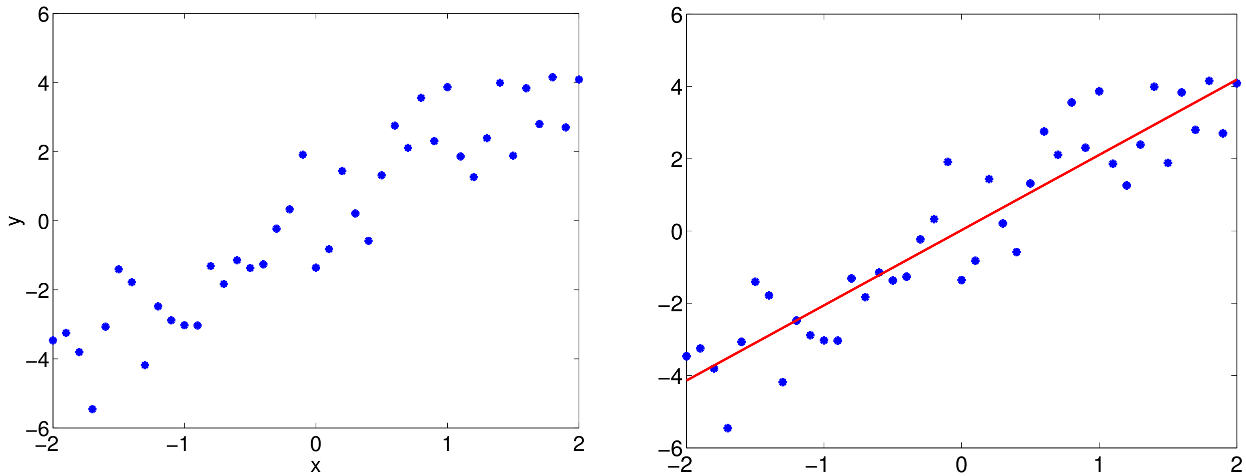
A generic regression problem

- The input attributes are given as fixed length vectors $\mathbf{x} = [x_1, \ldots, x_d]^T$, where each component such as x_i may be discrete or real valued.
- The outputs are assumed to be real valued $y \in \mathcal{R}$ (the values of actual outputs such as prices may be more restricted)
- We have access to a set of n *training examples*, $D_n = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, sampled independently at random from some fixed but unknown distribution $P(\mathbf{x}, y)$
- The goal is to minimize the prediction error/loss on new examples $(\mathbf{x}, y)$ drawn at random from the same $P(\mathbf{x}, y)$. The loss may be, for example, the squared loss

$$\text{Loss}(y, \hat{y}) = (y - \hat{y})^2$$

where $\hat{y}$ denotes our prediction in response to $\mathbf{x}$.

Types of predictions: regression function

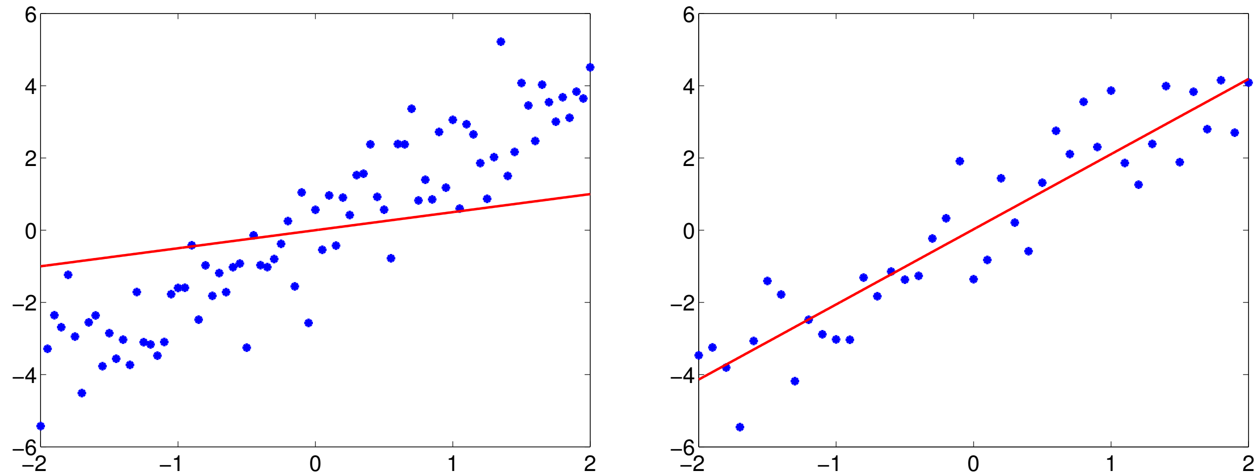


- We need to define a class of functions (types of predictions we will try to make) such as linear predictions

$$f(x; w_1, w_0) = w_0 + w_1 x$$

where w_1, w_0 are the *parameters* we need to set.

Estimation criterion



- In addition, we need a fitting/estimation criterion so as to be able to select appropriate values for the *parameters* w_1, w_0 based on the training set $D_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$.

  For example, we can use the *empirical loss*:

$$J_n(w_1, w_0) = \frac{1}{n} \sum_{t=1}^{n} (y_t - f(x_t; w_1, w_0))^2$$

(note: the loss here is the same as in evaluation)

Empirical loss: motivation

- Ideally, we would like to find the parameters w_1, w_0 that minimize the expected loss (unlimited training data):

$$J(w_1, w_0) = E_{(x,y)\sim P}\,(y - f(x; w_1, w_0))^2$$

where the expectation is over samples from $P(x, y)$.

- When the number of training examples n is large, however, the empirical error is approximately what we want

$$E_{(x,y)\sim P}\,(y - f(x; w_1, w_0))^2 \approx \frac{1}{n}\sum_{t=1}^{n}(y_t - f(x_t; w_1, w_0))^2$$

Linear regression: estimation

- We minimize the *empirical* squared loss

$$
\begin{aligned}
J_n(w_1, w_0) &= \frac{1}{n}\sum_{t=1}^{n} (y_t - f(x_t; w_1, w_0))^2 \\
&= \frac{1}{n}\sum_{t=1}^{n} (y_t - w_0 - w_1 x_t)^2
\end{aligned}
$$

By setting the derivatives with respect to w_1 and w_0 to zero we get necessary conditions for the "optimal" parameter values

$$
\begin{aligned}
\frac{\partial}{\partial w_1} J_n(w_1, w_0) &= 0 \\
\frac{\partial}{\partial w_0} J_n(w_1, w_0) &= 0
\end{aligned}
$$

Derivation

$$\frac{\partial}{\partial w_1} J_n(w_1, w_0) = \frac{\partial}{\partial w_1} \frac{1}{n} \sum_{t=1}^{n} (y_t - w_0 - w_1 x_t)^2$$

Derivation

$$
\begin{aligned}
\frac{\partial}{\partial w_1} J_n(w_1, w_0) &= \frac{\partial}{\partial w_1} \frac{1}{n} \sum_{t=1}^{n} (y_t - w_0 - w_1 x_t)^2 \\
&= \frac{1}{n} \sum_{t=1}^{n} \frac{\partial}{\partial w_1} (y_t - w_0 - w_1 x_t)^2
\end{aligned}
$$

Derivation

$$
\begin{aligned}
\frac{\partial}{\partial w_1} J_n(w_1, w_0) &= \frac{\partial}{\partial w_1} \frac{1}{n} \sum_{t=1}^{n} (y_t - w_0 - w_1 x_t)^2 \\
&= \frac{1}{n} \sum_{t=1}^{n} \frac{\partial}{\partial w_1} (y_t - w_0 - w_1 x_t)^2 \\
&= \frac{2}{n} \sum_{t=1}^{n} (y_t - w_0 - w_1 x_t) \frac{\partial}{\partial w_1} (y_t - w_0 - w_1 x_t)
\end{aligned}
$$

Derivation

$$\begin{aligned}
\frac{\partial}{\partial w_1} J_n(w_1, w_0) &= \frac{\partial}{\partial w_1} \frac{1}{n} \sum_{t=1}^{n} (y_t - w_0 - w_1 x_t)^2 \\
&= \frac{1}{n} \sum_{t=1}^{n} \frac{\partial}{\partial w_1} (y_t - w_0 - w_1 x_t)^2 \\
&= \frac{2}{n} \sum_{t=1}^{n} (y_t - w_0 - w_1 x_t) \frac{\partial}{\partial w_1} (y_t - w_0 - w_1 x_t) \\
&= \frac{2}{n} \sum_{t=1}^{n} (y_t - w_0 - w_1 x_t)(-x_t) = 0
\end{aligned}$$

Derivation

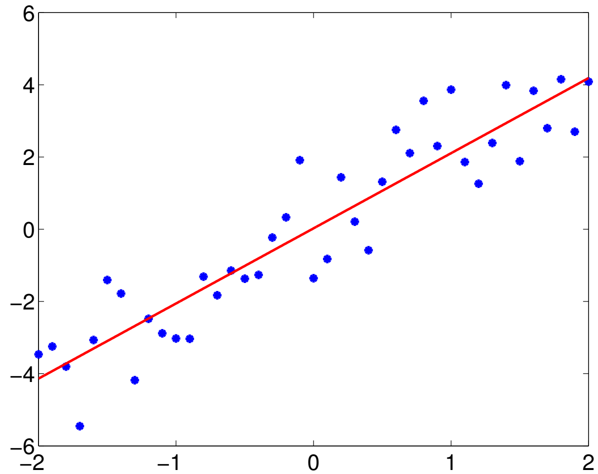
$$
\begin{aligned}
\frac{\partial}{\partial w_1} J_n(w_1, w_0) &= \frac{\partial}{\partial w_1} \frac{1}{n} \sum_{t=1}^{n} (y_t - w_0 - w_1 x_t)^2 \\
&= \frac{1}{n} \sum_{t=1}^{n} \frac{\partial}{\partial w_1} (y_t - w_0 - w_1 x_t)^2 \\
&= \frac{2}{n} \sum_{t=1}^{n} (y_t - w_0 - w_1 x_t) \frac{\partial}{\partial w_1} (y_t - w_0 - w_1 x_t) \\
&= \frac{2}{n} \sum_{t=1}^{n} (y_t - w_0 - w_1 x_t)(-x_t) = 0 \\
\frac{\partial}{\partial w_0} J_n(w_1, w_0) &= \frac{2}{n} \sum_{t=1}^{n} (y_t - w_0 - w_1 x_t)(-1) = 0
\end{aligned}
$$

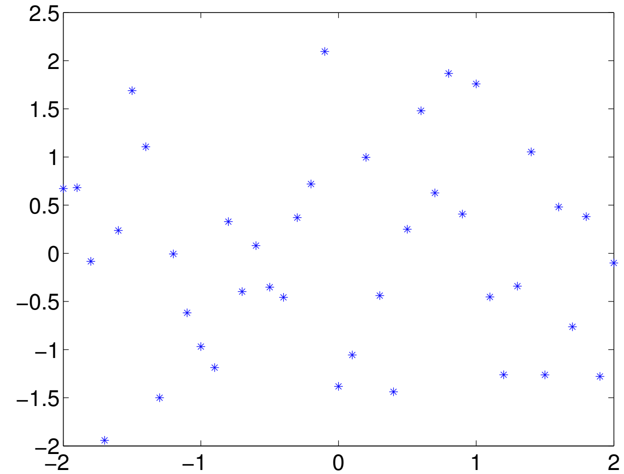
Interpretation

- The optimality conditions

$$\begin{aligned} \frac{2}{n}\sum_{t=1}^{n}(y_t - w_0 - w_1 x_t)(-x_t) &= 0 \\ \frac{2}{n}\sum_{t=1}^{n}(y_t - w_0 - w_1 x_t)(-1) &= 0 \end{aligned}$$

ensure that the prediction error $\epsilon_t = (y_t - w_0 - w_1 x_t)$ is decorrelated with any linear function of the inputs





Linear regression: matrix notation

- We can express the solution a bit more generally by resorting to a matrix notation

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

so that

$$\begin{aligned} \frac{1}{n}\sum_{t=1}^{n}(y_t - w_0 - w_1 x_t)^2 &= \frac{1}{n}\left\| \begin{bmatrix} y_1 \\ \cdots \\ y_n \end{bmatrix} - \begin{bmatrix} 1 & x_1 \\ \cdots & \cdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \right\|^2 \\ &= \frac{1}{n}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 \end{aligned}$$

Linear regression: solution

By setting the derivatives of $\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2/n$ to zero, we get the same optimality conditions as before, now expressed in a matrix form

$$
\begin{aligned}
\frac{\partial}{\partial \mathbf{w}} \frac{1}{n} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 &= \frac{\partial}{\partial \mathbf{w}} \frac{1}{n} (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w}) \\
&\cdots \\
&= \frac{2}{n} (\mathbf{X}^T \mathbf{y} - \mathbf{X}^T \mathbf{X}\mathbf{w}) = \mathbf{0}
\end{aligned}
$$

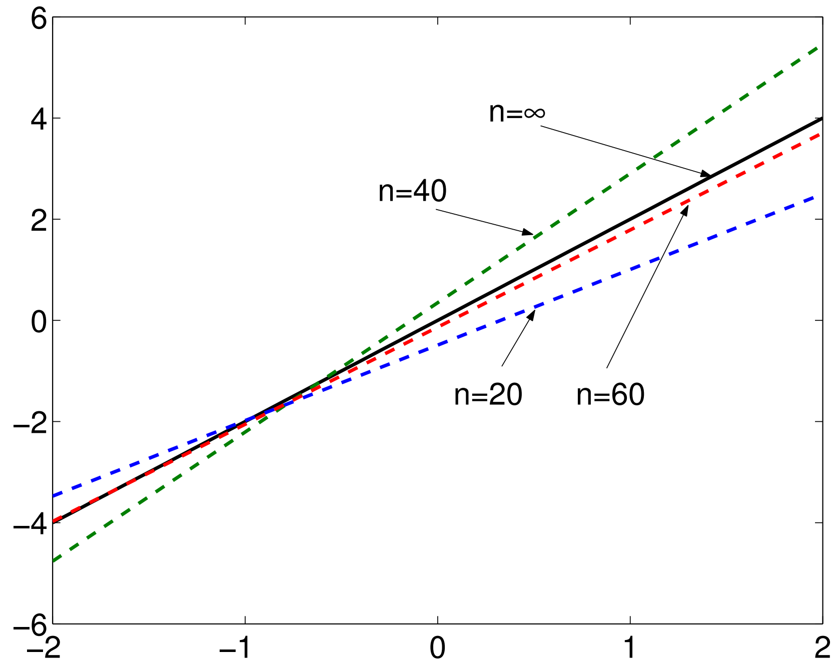
which gives

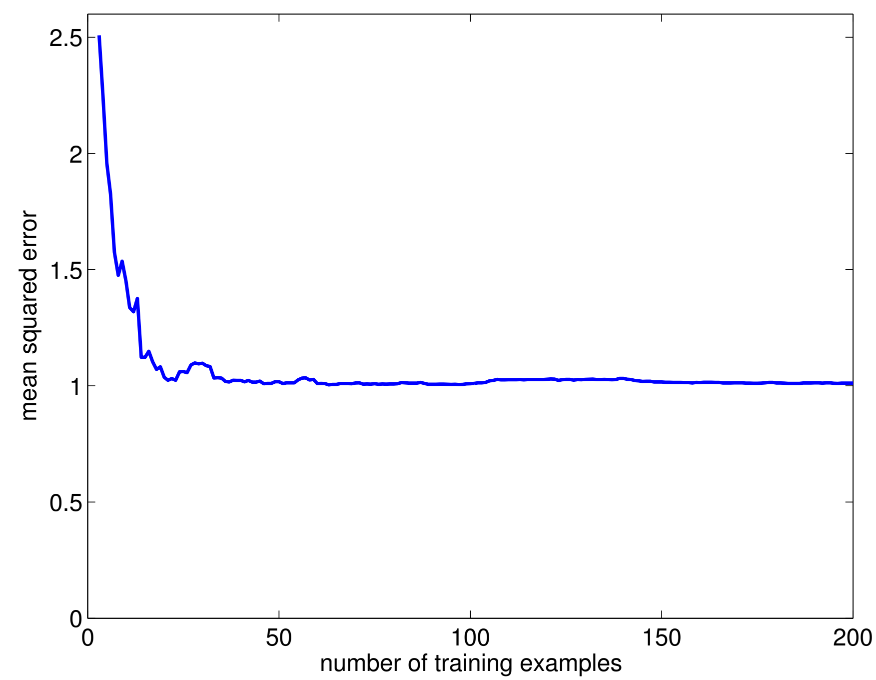
$$
\hat{\mathbf{w}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}
$$

- The solution is a linear function of the outputs y

Linear regression: generalization

- As the number of training examples increases our solution gets "better"





We'd like to understand the error a bit better

Linear regression: types of errors

- Structural error measures the error introduced by the limited function class (infinite training data):

$$\min_{w_1, w_0} E_{(x,y)\sim P} \left(y - w_0 - w_1 x\right)^2 = E_{(x,y)\sim P} \left(y - w_0^* - w_1^* x\right)^2$$

where (w_0^*, w_1^*) are the optimal linear regression parameters.

Linear regression: types of errors

- Structural error measures the error introduced by the limited function class (infinite training data):

$$\min_{w_1, w_0} E_{(x,y)\sim P} \left(y - w_0 - w_1 x\right)^2 = E_{(x,y)\sim P} \left(y - w_0^* - w_1^* x\right)^2$$

where (w_0^*, w_1^*) are the optimal linear regression parameters.

- Approximation error measures how close we can get to the optimal linear predictions with limited training data:

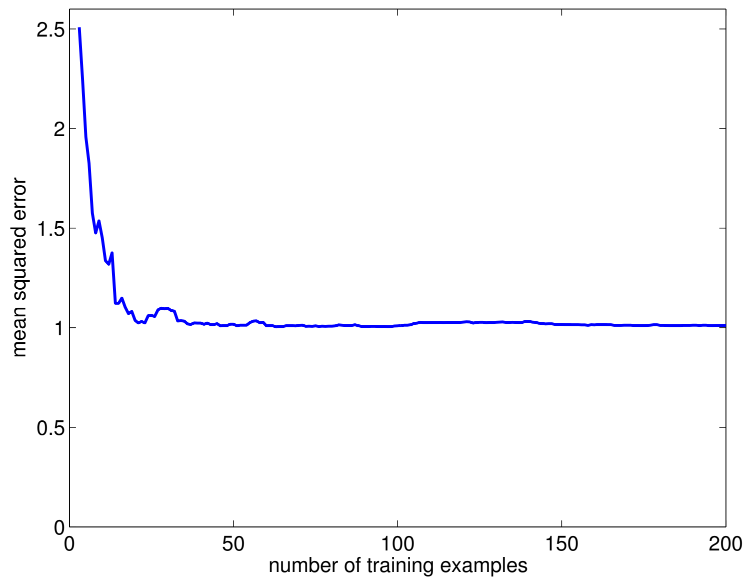
$$E_{(x,y)\sim P} \left(w_0^* + w_1^* x - \hat{w}_0 - \hat{w}_1 x\right)^2$$

where $(\hat{w}_0, \hat{w}_1)$ are the parameter estimates based on a small training set (therefore themselves random variables).

Linear regression: error decomposition

- The expected error of our linear regression function decomposes into the sum of structural and approximation errors

$$
\begin{aligned}
E_{(x,y)\sim P}\left(y - \hat{w}_0 - \hat{w}_1 x\right)^2 = \\
& E_{(x,y)\sim P}\left(y - w_0^* - w_1^* x\right)^2 + \\
& E_{(x,y)\sim P}\left(w_0^* + w_1^* x - \hat{w}_0 - \hat{w}_1 x\right)^2
\end{aligned}
$$



Error decomposition: derivation

$$
\begin{aligned}
& E_{(x,y)\sim P}\,(y - \hat{w}_0 - \hat{w}_1 x)^2 \\
&= E_{(x,y)\sim P}\left((y - w_0^* - w_1^* x) + (w_0^* + w_1^* x - \hat{w}_0 - \hat{w}_1 x)\right)^2 \\
&= E_{(x,y)\sim P}\,(y - w_0^* - w_1^* x)^2 \\
&\quad + E_{(x,y)\sim P}\,2(y - w_0^* - w_1^* x)(w_0^* + w_1^* x - \hat{w}_0 - \hat{w}_1 x) \\
&\quad + E_{(x,y)\sim P}\,(w_0^* + w_1^* x - \hat{w}_0 - \hat{w}_1 x)^2
\end{aligned}
$$

The second term has to be zero since the error $(y - w_0^* - w_1^* x)$ of the best linear predictor is necessarily decorrelated with any linear function of the input including $(w_0^* + w_1^* x - \hat{w}_0 - \hat{w}_1 x)$