Machine learning: lecture 2

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

Topics

- Brief review of background
- Linear regression
  - estimation criterion
  - least squares solution, properties
  - generalization, overfitting

Brief review of background

- Expectation and sample mean

$$E_{X \sim P}\{ X \} \;=\; E\{ X \} \approx \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

  where each x_i is a sample from P.

- Variance and sample variance

$$Var\{X\} \;=\; E\left\{(X - E\{X\})^2\right\} \approx \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Brief review of background cont'd

- Covariance and sample covariance

$$
\begin{aligned}
Cov\{X_1, X_2\} &= E\Big\{ (X_1 - E\{X_1\})(X_2 - E\{X_2\}) \Big\} \\
&\approx \frac{1}{n} \sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)
\end{aligned}
$$

where (x_{1i}, x_{2i}) is the i^{th} joint sample from P.

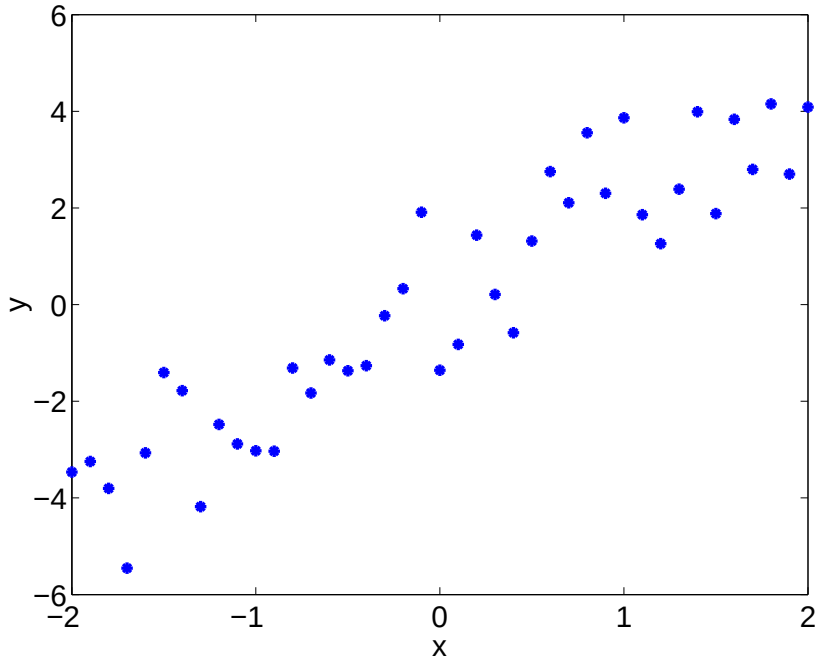
Brief review of background cont'd

- Conditional expectation:

$$E\{\, Y|x \,\} = \int_y y\, p(y|x) dy$$
$$E\{\, E\{\, Y|X \,\}\,\} = E\{\, Y \,\}$$

  where X and Y are random variables governed by a distribution p; x is a possible value of X.

Regression



- The goal is to predict responses/outputs for inputs
- We need to define
  1. *function class*, the type of predictions we consider
  2. *fitting criterion* (loss) that measures the degree of fit to the data

Linear regression

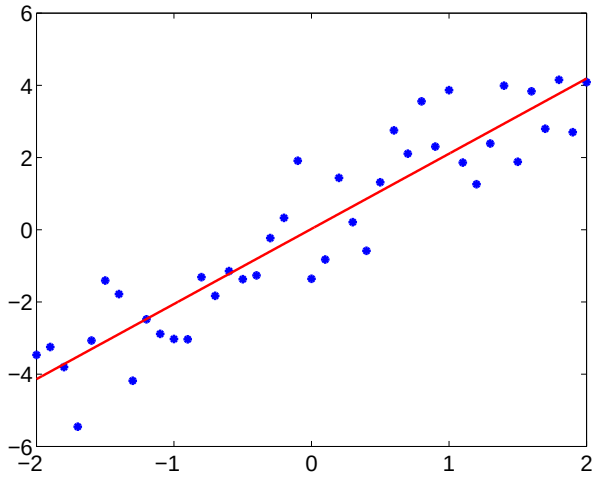
- Linear functions of one variable (two parameters)

$$f(x; \mathbf{w}) = w_0 + w_1 x, \quad \mathbf{w} = [w_0\, w_1]^T$$

  and a squared loss: $\mathrm{Loss}(y, f(x; \mathbf{w})) = (y - f(x; \mathbf{w}))^2/2$.
- Estimation based on minimizing the *empirical loss*

$$J_n(\mathbf{w}) = \sum_{i=1}^{n} \mathrm{Loss}(y_i, f(x_i; \mathbf{w}))$$

  with respect to the parameters $\mathbf{w}$.



Linear regression: estimation

- We minimize the *empirical* squared loss

$$J_n(\mathbf{w}) = \sum_{i=1}^{n} \text{Loss}(y_i, f(x_i; \mathbf{w})) = \sum_{i=1}^{n} (y_i - w_0 - w_1 x_i)^2/2$$

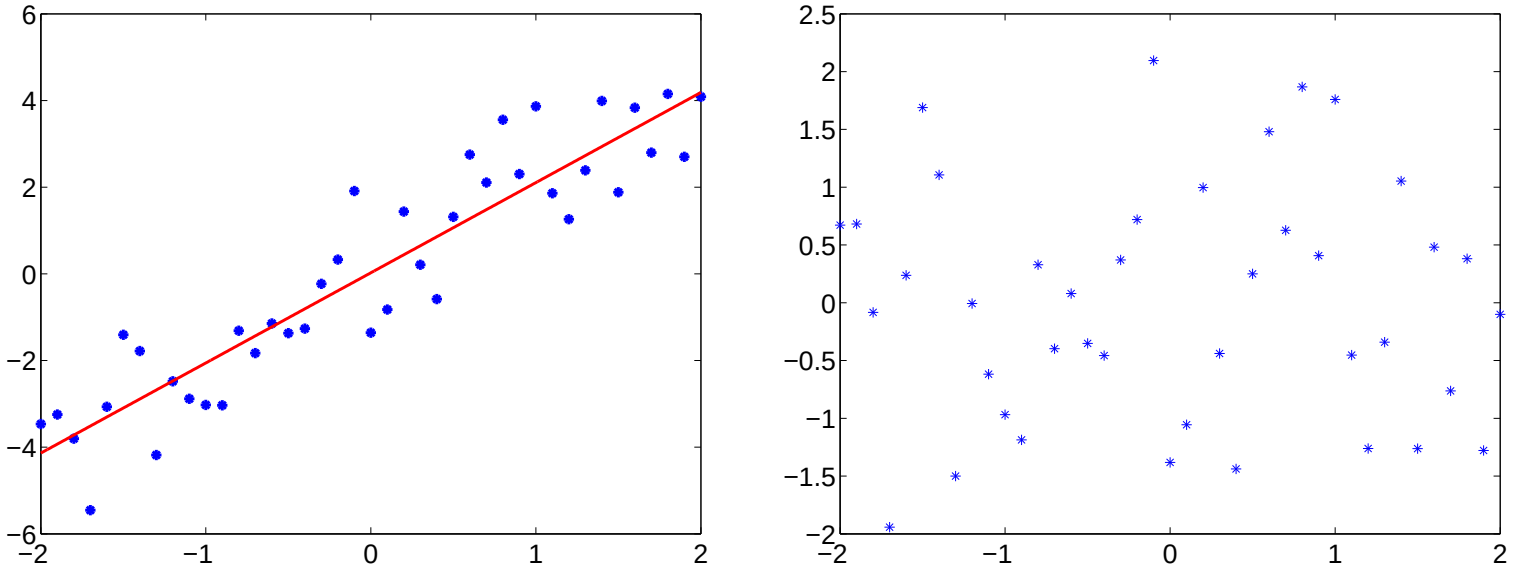
Setting the derivatives with respect to w_0 and w_1 to zero we get necessary conditions for the "optimal" parameter values

$$\begin{aligned} \frac{\partial}{\partial w_0} J_n(\mathbf{w}) &= -\sum_{i=1}^{n} (y_i - w_0 - w_1 x_i) = 0 \\ \frac{\partial}{\partial w_1} J_n(\mathbf{w}) &= -\sum_{i=1}^{n} (y_i - w_0 - w_1 x_i)\, x_i = 0 \end{aligned}$$

Note: These conditions mean that the prediction error $(y_i - w_0 - w_1 x_i)$ has zero mean and is decorrelated with the inputs x_i

Linear regression: estimation

- The prediction error $(y_i - w_0 - w_1 x_i)$ is decorrelated with the inputs x_i



Linear regression: estimation

$$\frac{\partial}{\partial w_0} J_n(\mathbf{w}) = -\sum_{i=1}^{n} (y_i - w_0 - w_1 x_i) = 0$$

$$\frac{\partial}{\partial w_1} J_n(\mathbf{w}) = -\sum_{i=1}^{n} (y_i - w_0 - w_1 x_i)\, x_i = 0$$

- Solution via matrix inversion

$$\begin{array}{rcrcl} w_0 \left(\sum_{i=1}^n 1\right) & + & w_1 \left(\sum_{i=1}^n x_i\right) & = & \sum_{i=1}^n y_i \\ w_0 \left(\sum_{i=1}^n x_i\right) & + & w_1 \left(\sum_{i=1}^n x_i^2\right) & = & \sum_{i=1}^n y_i x_i \end{array}$$

  or $\Phi \mathbf{w} = b$, where

$$\Phi = \begin{bmatrix} \sum_{i=1}^n 1 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}, \quad b = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n y_i x_i \end{bmatrix}$$

- If Φ is invertible, we get our parameter estimates via $\hat{\mathbf{w}} = \Phi^{-1} b$

Linear regression

- In a matrix notation, we minimize:

$$\frac{1}{2}\left\| \begin{bmatrix} y_1 \\ \dots \\ y_n \end{bmatrix} - \begin{bmatrix} 1 & x_1 \\ \dots & \dots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \right\|^2$$

or $\frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2$

By setting the derivatives to zero, we get

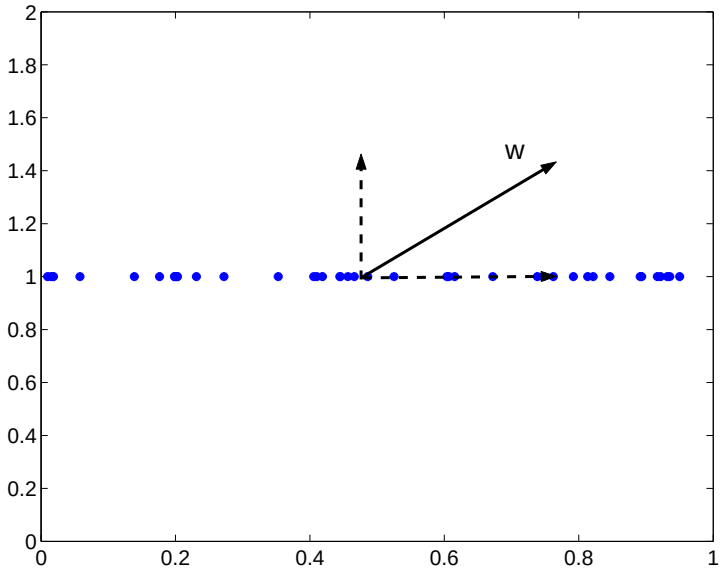
$$\mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{X}\mathbf{w} = 0 \;\; \Rightarrow \;\; \hat{\mathbf{w}} = \underbrace{(\mathbf{X}^T\mathbf{X})}_{\Phi}^{-1} \underbrace{\mathbf{X}^T\mathbf{y}}_{b}$$

Note: the solution is a linear function of the outputs y

Linear regression: pseudo-inverse

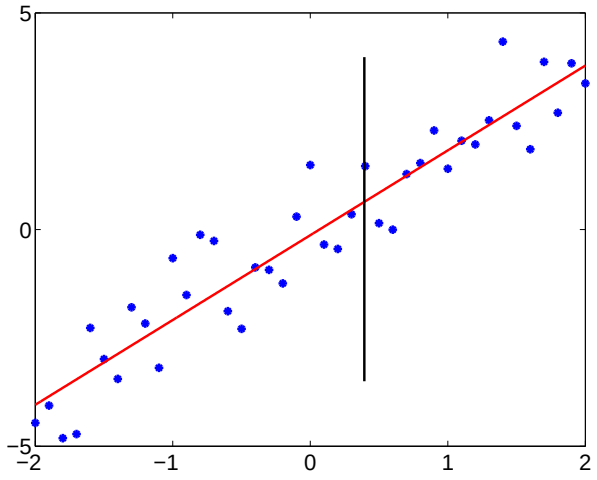
- 2-D example: $\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^\dagger \mathbf{X}^T\mathbf{y}$

$$y_i \approx f(\mathbf{x}_i; \hat{\mathbf{w}}) = \hat{w}_0 + \hat{w}_1 x_{1i} + \hat{w}_2 x_{2i} = \hat{\mathbf{w}}^T \begin{bmatrix} 1 \\ x_{1i} \\ x_{2i} \end{bmatrix}$$



- We find the solution in the subspace spanned by the examples (weight vector set to zero in the orthogonal dimensions)

Properties of estimates



- Suppose the mean response (output) for any input x can indeed be modeled as a linear function with some "true" parameters $\mathbf{w}^*$:

$$E\{\, y|x \,\} = f(x; \mathbf{w}^*) = \mathbf{w}^{*T} \begin{bmatrix} 1 \\ x \end{bmatrix}$$

  We can ask if our estimate $\hat{\mathbf{w}}$ (based on a limited training set) is in any sense close to $\mathbf{w}^*$.

Bias

- One measure of deviation is *bias*, which gauges any systematic deviation from $\mathbf{w}^*$:

$$\text{Bias} = E\{\, \hat{\mathbf{w}} \,\} - \mathbf{w}^*$$

  where the expectation is taken with respect to resampled training sets of the same size; each input/output pair (x, y) in the training set is assumed to be an independent sample from some distribution P

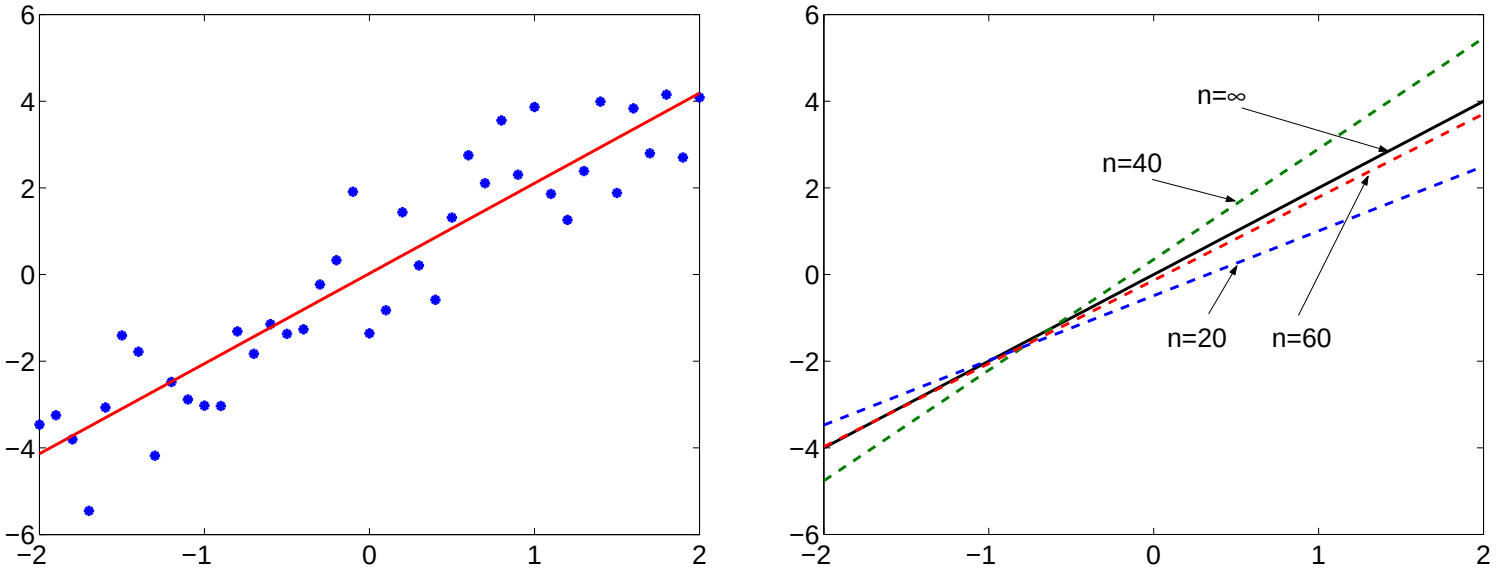
- In linear regression the estimate $\hat{\mathbf{w}}$ is *unbiased*, i.e., $E\{\, \hat{\mathbf{w}} \,\} - \mathbf{w}^* = 0$ (problem set)

- This means that the predictions are unbiased as well:

$$E\{\, f(x; \hat{\mathbf{w}}) \,\} = E\left\{ \hat{\mathbf{w}}^T \begin{bmatrix} 1 \\ x \end{bmatrix} \right\} = \mathbf{w}^{*T} \begin{bmatrix} 1 \\ x \end{bmatrix} = f(x; \mathbf{w}^*)$$

Linear regression: generalization

- We'd like to understand how our linear predictions "improve" as a function of the number of training examples $\{(x_1, y_1), \ldots, (x_n, y_n)\}$



We assume that there is a systematic relation between x and y: each training example (x, y) is an *independent* sample from a fixed but unknown distribution P

Linear regression: generalization

Training examples $\{(x_1, y_1), \ldots, (x_n, y_n)\}$

Test examples $\{(x_{n+1}, y_{n+1}), \ldots, (x_{n+N}, y_{n+N})\}$

$\hat{\mathbf{w}}$ is the parameter estimate from the training examples.

- Types of errors:

$$\text{Mean training error} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{w}_0 - \hat{w}_1 x_i)^2$$

$$\text{Mean test error} = \frac{1}{N} \sum_{i=n+1}^{n+N} (y_i - \hat{w}_0 - \hat{w}_1 x_i)^2$$

$$\text{“Generalization” error} = E_{(x,y)\sim P} \left\{ (y - \hat{w}_0 - \hat{w}_1 x)^2 \right\}$$

(note: $\hat{w}_0$ and $\hat{w}_1$ are themselves random variables as they depend on the training set)

Linear regression: generalization

- We can decompose the "generalization" error

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

into two terms:

1. error of the best predictor in the class

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

2. and how well we approximate the best predictor

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

- This holds for any input/output relation depicted by the distribution P

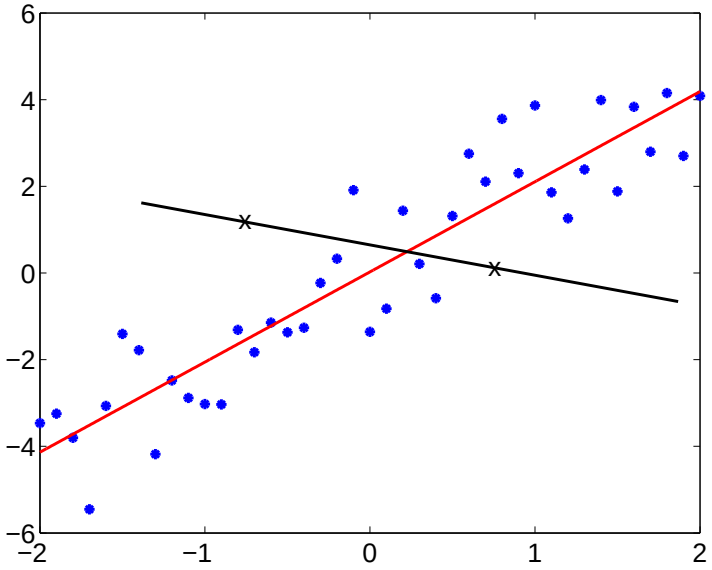
Brief derivation

$$
\begin{aligned}
(y - \hat{w}_0 - \hat{w}_1 x)^2 = & \\
& \big((y - (w_0^* + w_1^* x) + (w_0^* + w_1^* x) - (\hat{w}_0 + \hat{w}_1 x)\big)^2 \\
= \quad & \big(y - (w_0^* + w_1^* x)\big)^2 + \\
& +2\big(y - (w_0^* + w_1^* x)\big)\big((w_0^* + w_1^* x) - (\hat{w}_0 + \hat{w}_1 x)\big) + \\
& +\big((w_0^* + w_1^* x) - (\hat{w}_0 + \hat{w}_1 x)\big)^2
\end{aligned}
$$

The cross-terms (blue) vanish when we take expectation with respect to $(x, y) \sim P$. ($\mathbf{w}^*$ is the best linear predictor)

Overfitting

- With too few training examples our linear regression model may achieve zero training error but nevertless has a large generalization error



When the training error no longer bears any relation to the generalization error the model *overfits* the data