Thursday, 9/3/2020

# EE 660

# MACHINE LEARNING FROM SIGNALS: FOUNDATIONS AND METHODS

Prof. B. Keith Jenkins

**Lecture 4**

## Announcements

- Class projects and quizzes
  - There will be 1 end-of-semester quiz and no midterm quiz
  - There will be 2 projects - a (smaller) midterm project and a (larger) final project
  - The above were preferred by a large margin of students
- End-of-semester quiz
  - Tuesday, 11/24/2020, 5:30 - 7:00 PM
  - (will end earlier if the quiz is shorter than 90 min.)
  - This time has no remaining known conflicts
  - If you have an unmovable conflict, please email me asap.
- Final project
  - Will be due on Thur., 12/3/2020
- First homework will be posted this Friday
- From now on, most reading assignments will be given in each homework

## Today's Lecture

- Data notation
- Comment on definition of dataset D
  - y|x or y,x
- MLE Regression (part 2)
- Ridge regression

# Dataset Notation (1)

(2) $\mathcal{D} = \{\underline{x}_i, y_i\}_{i=1}^{N}$. Later: $\mathcal{D}_{Tr}$ with $N = N_{Tr}$ (training set)

$\mathcal{D}_{val}$ with $N = N_{val}$ (validation set)

$\mathcal{D}_{Test}$ with $N = N_{Test}$ (test set)

Design matrix: $\underline{\underline{X}} = \begin{bmatrix} \underline{x}_1^T \\ \underline{x}_2^T \\ \vdots \\ \underline{x}_N^T \end{bmatrix}_{N \times D}$ Vector of data-point labels $\Big\}$ $\underline{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$

We use $\mathcal{D}$ for both set notation as in (2), and matrix-vector pair as in:

$$\mathcal{D} = \underline{\underline{X}}, \underline{y} \quad \text{or} \quad \underline{y}, \underline{\underline{X}}$$

$$\therefore \quad p(\mathcal{D} / \underline{\theta}) = p(\underline{y}, \underline{\underline{X}} \,|\, \underline{\theta})$$

but what we often want is $p(\underline{y} \,|\, \underline{\underline{X}}, \underline{\theta})$ or $p(\underline{\underline{X}} \,|\, \underline{y}, \underline{\theta})$.

p.3

**EE 660** **Notes on** $p\left(\mathcal{D}\middle|\underline{\theta}\right)$, $p\left(\underline{y},\underline{\underline{X}}\middle|\underline{\theta}\right)$, $p\left(\underline{y}\middle|\underline{\underline{X}},\underline{\theta}\right)$ **Fall 2020**

Using probability relations we have:

$$p\left(y,\underline{x}\middle|\underline{\theta}\right)=p\left(y\middle|\underline{x},\underline{\theta}\right)p\left(\underline{x}\middle|\underline{\theta}\right)=p\left(y\middle|\underline{x},\underline{\theta}\right)p\left(\underline{x}\right)$$

where for the last step we have dropped the last condition on $\underline{\theta}$ because it tells us nothing useful about $p\left(\underline{x}\right)$. If we are interested in maximizing (or minimizing) the likelihood, we will take:

$$\begin{aligned}
\underset{\underline{\theta}}{\arg\max}\, p\left(\mathcal{D}\middle|\underline{\theta}\right) &= \underset{\underline{\theta}}{\arg\max}\left\{\prod_{i=1}^{N} p\left(y_i,\underline{x}_i\middle|\underline{\theta}\right)\right\} \\
&= \underset{\underline{\theta}}{\arg\max}\left\{\prod_{i=1}^{N} p\left(y_i\middle|\underline{x}_i,\underline{\theta}\right)p\left(\underline{x}_i\right)\right\} = \underset{\underline{\theta}}{\arg\max}\left\{\underbrace{\cancel{\left(\prod_{i=1}^{N} p\left(\underline{x}_i\right)\right)}}_{\text{Constant of } \underline{\theta} \; >0}\prod_{i=1}^{N} p\left(y_i\middle|\underline{x}_i,\underline{\theta}\right)\right\} \\
&= \underset{\underline{\theta}}{\arg\max}\left\{\prod_{i=1}^{N} p\left(y_i\middle|\underline{x}_i,\underline{\theta}\right)\right\}
\end{aligned}$$

and to obtain the last line, $\prod_{i=1}^{N} p\left(\underline{x}_i\right)$ was dropped because it is a positive multiplicative term that is a constant of $\underline{\theta}$. This can equivalently be seen by using the log likelihood instead:

$$\begin{aligned}
\underset{\underline{\theta}}{\arg\max}\, p\left(\mathcal{D}\middle|\underline{\theta}\right) &= \underset{\underline{\theta}}{\arg\max}\left\{\ln\prod_{i=1}^{N} p\left(y_i,\underline{x}_i\middle|\underline{\theta}\right)\right\} \\
&= \underset{\underline{\theta}}{\arg\max}\left\{\ln\prod_{i=1}^{N} p\left(y_i\middle|\underline{x}_i,\underline{\theta}\right)p\left(\underline{x}_i\right)\right\} \\
&= \underset{\underline{\theta}}{\arg\max}\left\{\sum_{i=1}^{N}\ln\left[p\left(y_i\middle|\underline{x}_i,\underline{\theta}\right)p\left(\underline{x}_i\right)\right]\right\} \\
&= \underset{\underline{\theta}}{\arg\max}\left\{\sum_{i=1}^{N}\left[\ln p\left(y_i\middle|\underline{x}_i,\underline{\theta}\right)+\ln p\left(\underline{x}_i\right)\right]\right\} \\
&= \underset{\underline{\theta}}{\arg\max}\left\{\sum_{i=1}^{N}\ln p\left(y_i\middle|\underline{x}_i,\underline{\theta}\right)\right\}
\end{aligned}$$

and to obtain the last line, the additive terms that don't depend on $\underline{\theta}$ have been dropped.

So, when the goal of using the likelihood is to find its argmax or argmin w.r.t. $\underline{\theta}$, we can replace $p\left(y_i,\underline{x}_i\middle|\underline{\theta}\right)$ directly with $p\left(y_i\middle|\underline{x}_i,\underline{\theta}\right)$.

## 2. Objective function

$$J_1(\underline{w}, \mathcal{D}) = ? = -\ln p(\mathcal{D} | \underline{w}) = NLL(\underline{w})$$

(or $-p(\mathcal{D}|\underline{w})$)

$$\ln p(\mathcal{D}|\underline{w}) = \sum_{i=1}^{N} \ln p(y_i | \underline{x}_i, \underline{w}_i) \qquad \text{(i.d.)}$$

$$= \sum_{i=1}^{N} \ln N(y_i | \underline{w}^T \underline{x}_i, \sigma^2) \qquad \text{(Gaussian model)}$$

Can re-write $J_1$ as: [M Eq. 7.5-7.9]

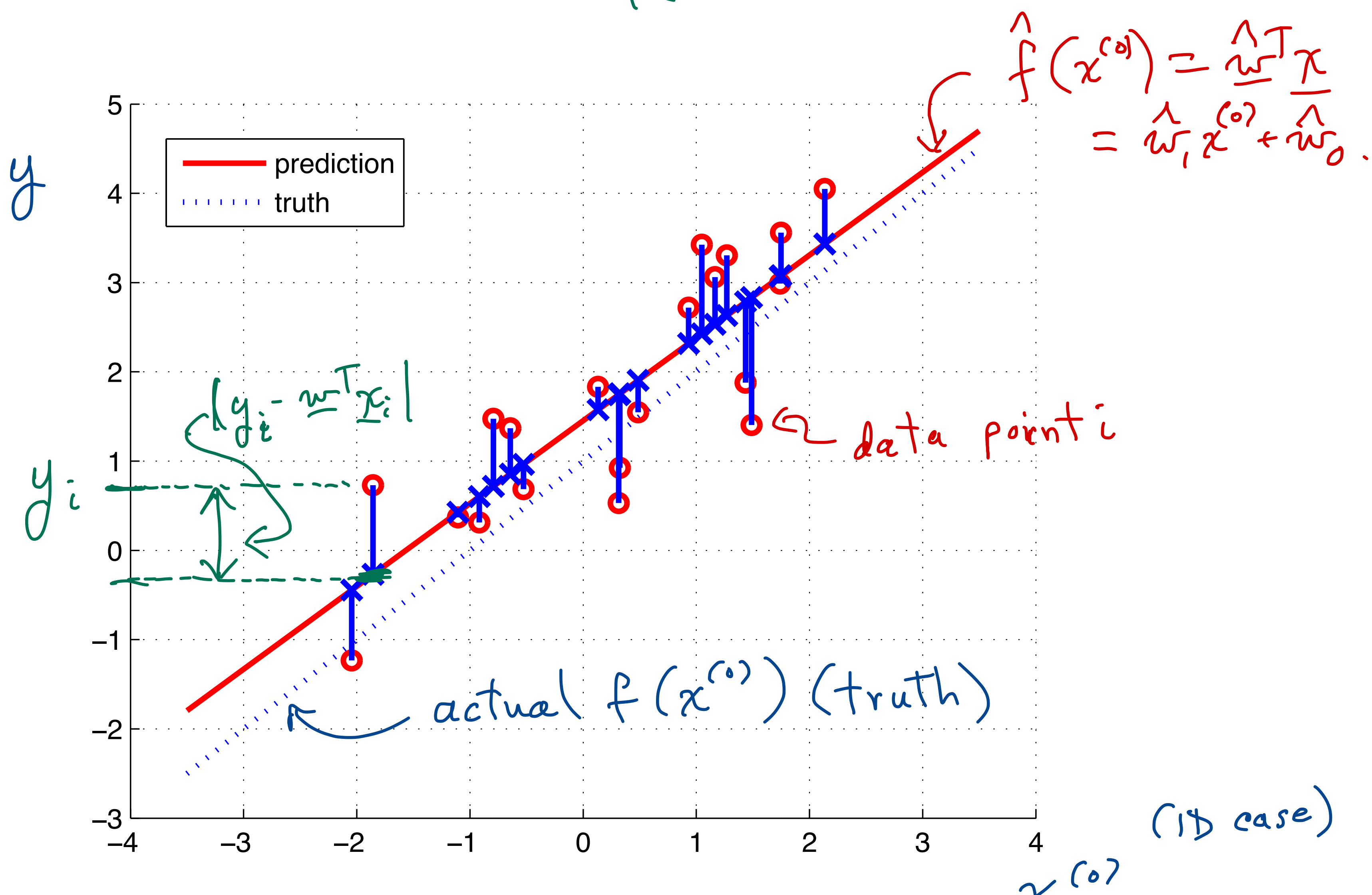
$$J_1(\underline{w}, \mathcal{D}) = \underbrace{\frac{1}{2\sigma^2}}_{\text{multiplicative constant} > 0} \sum_{i=1}^{N} (y_i - \underline{w}^T \underline{x}_i)^2 + \underbrace{\frac{N}{2} \ln(2\pi\sigma^2)}_{\text{constant of } \underline{w}}$$

Simplify:

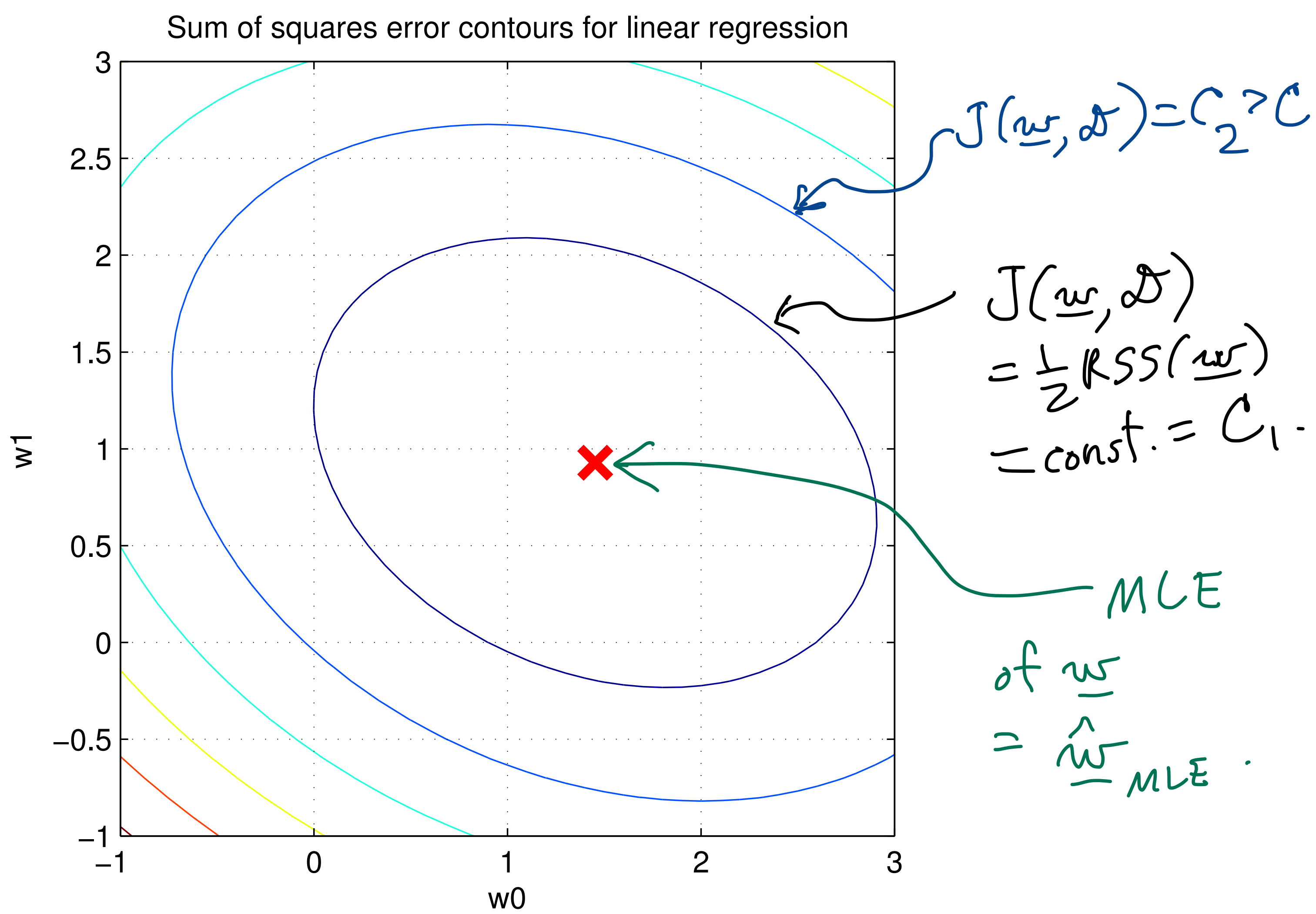
$$\Rightarrow \boxed{\text{Let } J(\underline{w}, \mathcal{D}) = \frac{1}{2} RSS(\underline{w}) = \frac{1}{2} \sum_{i=1}^{N} (y_i - \underline{w}^T \underline{x}_i)^2 = \frac{1}{2} \| \underline{y} - \underline{\underline{X}}\, \underline{w} \|_2^2}$$

$$\hat{y} \sim p(\hat{y} \mid \underline{x}, \underline{w}) = N(\hat{y} \mid \underbrace{\underline{w}^T \underline{x}}, \sigma^2)$$

$$\hat{f}(x) = \underline{\hat{w}}^T \underline{x}$$




Murphy Fig. 7.2(a)




Murphy Fig. 7.2(b)

# 3. Optimization method

Which method ?

- Gradient descent (stochastic or batch ...)
  - → neural-network approach
  - → very large dataset

- Solving $\nabla_{\underline{w}} J(\underline{w}, \mathcal{D}) = \underline{0}$ algebraically.
  - → Pseudoinverse solution.
  - → Non-neural approaches

→ Solving gives $\hat{\underline{w}}$

$$\underline{\underline{X}}^T \underline{\underline{X}} \, \hat{\underline{w}} = \underline{\underline{X}}^T \underline{y}$$

if $(\underline{\underline{X}}^T \underline{\underline{X}})$ is invertable, then:

$$\boxed{\hat{\underline{w}}_{OLS} = \underline{\underline{X}}^{-} \underline{y} = (\underline{\underline{X}}^T \underline{\underline{X}})^{-1} \underline{\underline{X}}^T \underline{y}}$$

= ordinary least squares solution.
= pseudoinv. sol'n.

$\hat{\underline{w}}_{MLE} \longleftrightarrow \hat{\underline{w}}_{OLS}$ ?

general MLE
(any $p(y \mid \underline{x}, \underline{w})$)

Normal case
$p(\cdot) = N(y \mid \underline{w}^T \underline{x}, \sigma^2)$

# Ridge Regression [Murphy 7.5]

$(\underline{w} = \underline{w}^{(0)})$

→ Use MAP (maximum a posteriori) estimation of $\underline{w}$.

Which is the MAP est. of $\underline{\theta}$?

(a) $\hat{\underline{\theta}} = \underset{\underline{\theta}}{\operatorname{argmax}}\ p(\mathcal{D}|\underline{\theta})$

(b) $\hat{\underline{\theta}} = \underset{\underline{\theta}}{\operatorname{argmax}}\ p(\underline{\theta}|\mathcal{D})$ ←

Model is same as MLE regression:

$$\begin{cases} y \sim p(y|\underline{x}, \underline{\theta}) \\ \quad = N(y|\underline{w}^T\underline{x}, \sigma^2) & \text{— Linear, Gaussian} \\ \text{or} \quad = N(y|\underline{w}^T\underline{\phi}(\underline{x}), \sigma^2) & \text{— Nonlinear, Gaussian} \end{cases}$$

Hypothesis set (linear, Gaussian case with $\sigma^2$ given):

$$\mathcal{H} = \{\hat{y}(\underline{x}) \sim N(\hat{y}|\underline{w}^T\underline{x}, \sigma^2) \mid \underline{w} \in \mathbb{R}^{D+1},\ D \in \mathbb{Z}^{\geq 0}\}$$

$\mathbb{Z}^{\geq 0}$ = set of positive integers

Is $\mathcal{H}$ the same as for MLE regression (linear, Gaussian case, $\sigma^2$ given)?

→ Yes.

What's different here? → objective function

$$\hat{\underline{\theta}} = \underset{\underline{\theta}}{\operatorname{argmax}}\; p(\underline{\theta} \mid \mathcal{D}) = \underset{\underline{\theta}}{\operatorname{argmax}} \left\{ \frac{p(\mathcal{D} \mid \underline{\theta})\, p(\underline{\theta})}{\underbrace{p(\mathcal{D})}_{\text{const. of } \underline{\theta}}} \right\}$$

$$= \underset{\underline{\theta}}{\operatorname{argmax}} \{ \underbrace{p(\mathcal{D} / \underline{\theta})\, p(\underline{\theta})}_{\text{always} \geq 0.} \}$$

(1) $$\hat{\underline{\theta}} = \underset{\underline{\theta}}{\operatorname{argmax}} \{ \ln \underbrace{p(\mathcal{D} \mid \underline{\theta})}_{\substack{\text{likelihood} \\ \text{of } \underline{\theta} \\ \text{(same as MLE)}}} + \ln \underbrace{p(\underline{\theta})}_{\text{prior for } \underline{\theta}} \}$$

(2) $$\ln p(\mathcal{D} \mid \underline{\theta}) = \sum_{i=1}^{N} \ln \mathcal{N}(y_i \mid w_0 + \underline{w}^T \underline{x}_i, \sigma^2)$$