Thursday, 8/27/2020

EE 660

MACHINE LEARNING FROM SIGNALS: FOUNDATIONS AND METHODS

Prof. B. Keith Jenkins

Lecture 2

p.1

Lecture 2 | EE 660 | Aug 27, 2020

Announcements

- Piazza is up and running; there is a link on D2L (navigation bar near the top)
  - Piazza is your primary source for questions and answers outside of classes and office hours.
- Discussion session: Fri., 3:30–4:20 PM
- One more TA:
  - Fernando Valladares Monteiro
    - fvallada@usc.edu
- Professor will hold informal office hours after class today (3:30–4:00 PM). Link is in D2L, "Virtual Meetings" in the navigation bar near the top.

Reading

- Regression, part 1
  - Murphy 1.4.5 (Linear Regression introduction)
  - Murphy 7.1 - 7.5, inclusive
  - Sections with asterisks (for example 7.4*) are optional; 7.3.2 is also optional (even though it has no asterisk).

Today's Lecture

- Course outline
- Regression and classification examples
- Key issues and concepts in ML

Example 1 – Classification

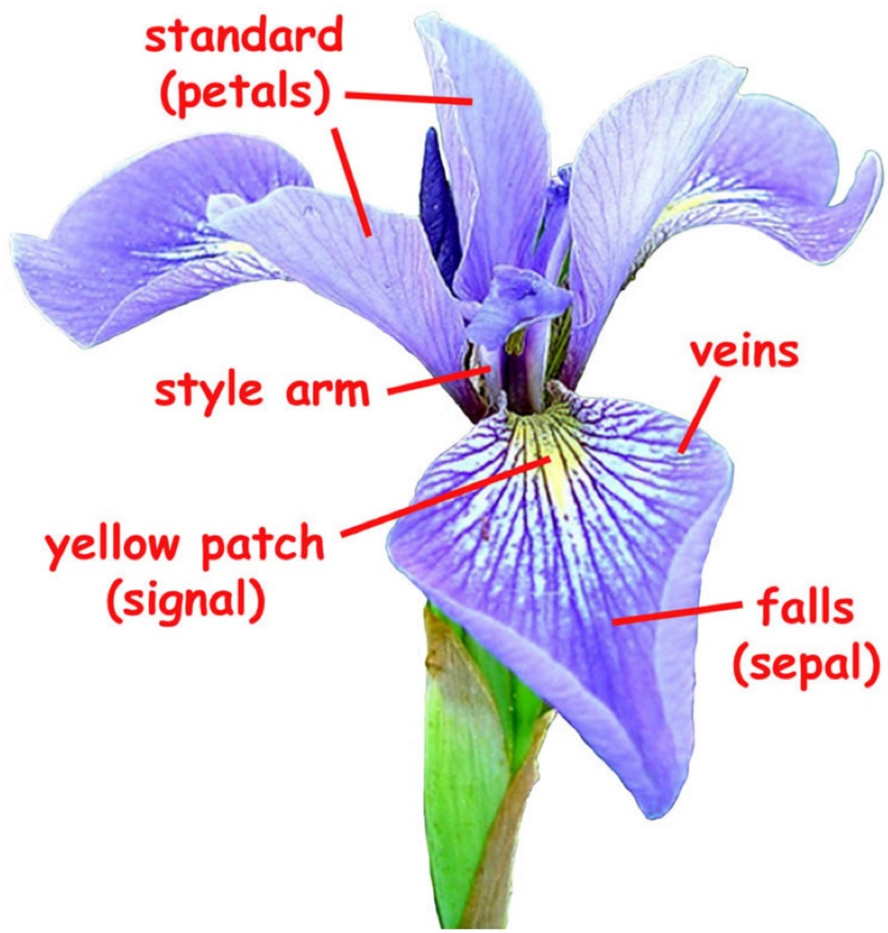
Flowers: Iris variety classification.

(a) setosa (b) versicolor (c) virginica

Murphy Fig. 1.3.
Iris flower types

Features to extract

$$\underline{x} = \begin{bmatrix} \text{petal length} \\ \text{petal width} \\ \text{sepal length} \\ \text{sepal width} \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$$



Iris parts. From USDA Forest Service at:
https://www.fs.fed.us/wildflowers/beauty/iris/flowers.shtml

Class assignment

For input (iris) $\underline{x}_i$

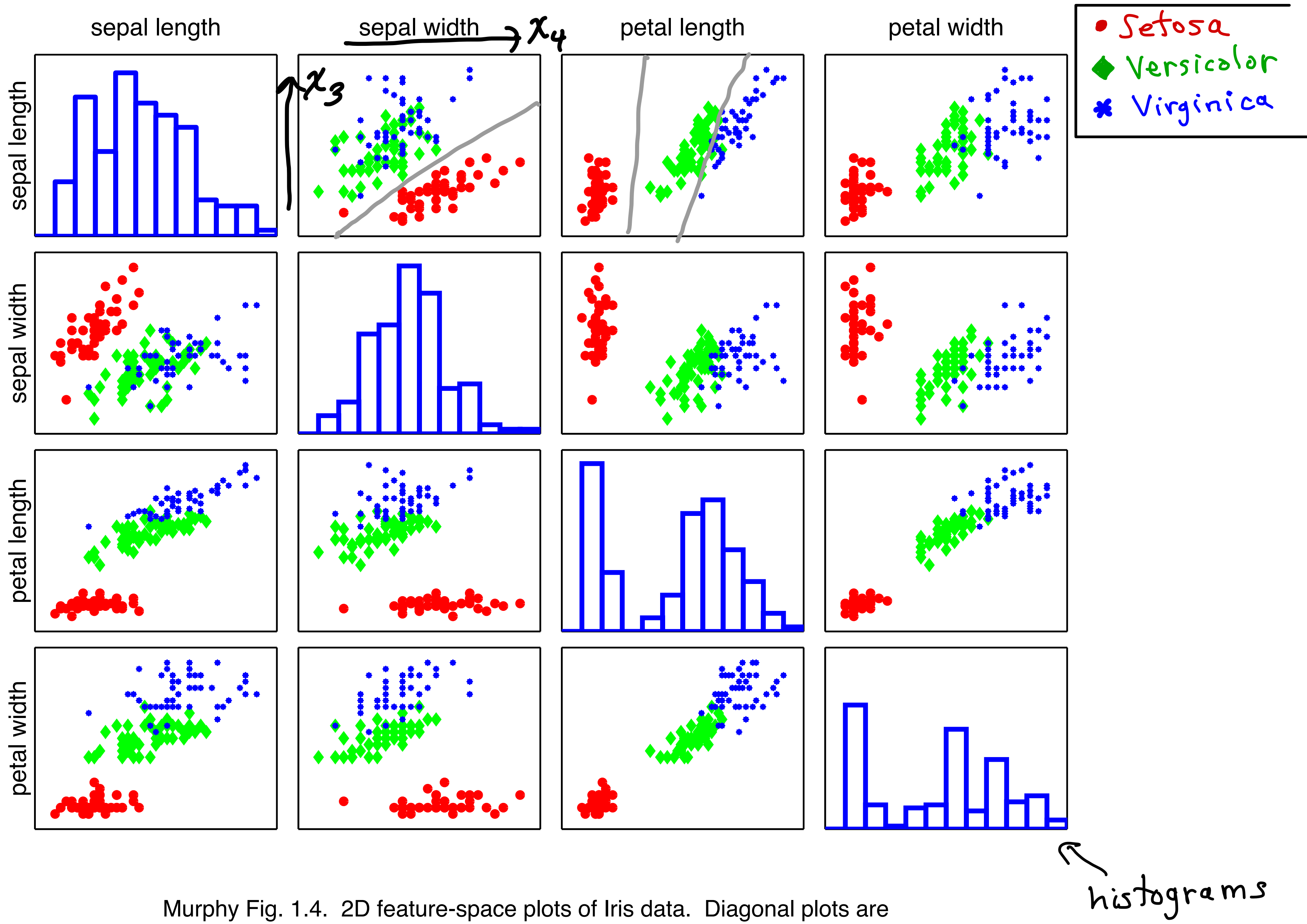
Output prediction $\hat{y}_i = \hat{y}(\underline{x}_i)$

= 1 of {setosa, versicolor, virginica}

or: $p(\hat{y}_i = c \mid \underline{x}_i)$

= posterior probability

Training data: $\mathcal{D}_{Tr} = \{(\underline{x}_i, y_i)\}_{i=1}^{N_{Tr}} \rightarrow$ plot it?



Murphy Fig. 1.4. 2D feature-space plots of Iris data. Diagonal plots are histograms.

Example 2 – Regression

Predict house price $\qquad y =$ (actual) price of house

input attributes (features): $\underline{x} = \begin{bmatrix} \text{Living area} \\ \text{Number of rooms} \\ \text{Age of house} \\ \text{Location 1} \\ \text{Location 2} \\ \vdots \end{bmatrix}$

output prediction:

$$\hat{y}_i = \hat{y}(\underline{x}_i) = \hat{f}(\underline{x}_i) = \text{estimate or prediction of house } i \text{ price}$$

also $\hat{y}(\underline{x}) = \hat{f}(\underline{x}) =$ prediction for any input values $\underline{x}$.

or $p(\hat{y}_i \mid \underline{x}_i) =$ posterior pdf of house i price prediction

Example 3 – Regression

Stock price prediction, for a given company.

Input data:

$r_1(t)$ = stock price at time t (from history)
$r_2(t)$ = sales reported for period ending at time t
$\vdots$

Output prediction:

$\hat{y}(\underline{x}_i)$ = stock price at future time t'

or $p(\hat{y}_i \mid \underline{x}_i, t')$ = pdf of stock price at future time t'

→ Can develop features from each signal $r_j(t)$ to define a feature space.

p-7

Course Outline

Number of lectures per topic is approximate.

Introduction

1. Course introduction [Murphy] {1 lecture}

   Administrative information; introduction to the course and to machine learning

2. Key issues and concepts in machine learning. {1 lecture}

Methods

Regression

3. Multidimensional regression [Murphy] {3 lectures}

   Linear regression, maximum-likelihood and MAP estimation, ridge regression, Bayesian regression. Learning linear and nonlinear relationships.

4. Logistic regression [Murphy] {1 lecture}

Foundations of learning: Bayesian

5. Bayesian concept learning {1 lecture}

Foundations

Foundations of learning: complexity

6. Feasibility of learning [AML] {1.5 lectures}

   Deterministic and statistical views; Hoeffding inequality (for bounding expected error on unlabeled data); inductive bias (model or data assumptions; e.g., parametric models, local smoothness)

7. Complexity of learning 1: generalization; estimation of error on new data; implications in dataset usage [AML] {3 lectures}

   Generalization bound, effective number of hypotheses, VC dimension, model complexity, sample complexity, dataset methodologies

8. Complexity of learning 2 [AML] {1.5 lectures}

   Bias-variance decomposition, learning curves, overfitting

Foundations & methods

Foundations and methods of learning: managing and controlling complexity

9. Regularization part 1 [AML] {1 lecture}

   Regularization as soft order constraints

10. Model selection [AML and Murphy] {1 lecture}

    Model selection and validation; consequences on generalization error bounds

11. Regularization part 2; feature reduction; sparsity [Murphy] {2 lectures}

    Bayesian and MAP estimation for feature reduction; quadratic regularization; l_1 regularization, lasso, and sparsity; comparison of l_1 and l_2 regularizers; nonconvex regularizers and l_0 regularization; bridge regression*

Foundations

12. Principles and pitfalls of learning [AML] {1 lecture}

    Occam's Razor, Axiom of Non-Falsifiability, Sampling Bias, Data Snooping

Methods

Graphical and nonlinear methods of learning

13. Boosting techniques and decision trees [Murphy] {3 lectures}

    Adaptive basis models; classification and regression trees (CART); random forests; boosting (Adaboost).

Methods

Semi-supervised and unsupervised learning methods

14. Semi-supervised learning for classification [Zhu] {3 lectures}

    Overview, including inductive vs. transductive semi-supervised learning; mixture models and Expectation Maximization for semi-supervised learning.

use much more data than is labeled.

15. Unsupervised learning for clustering: statistical techniques [Xu] {1 lecture}

    Statistical techniques including mixture models; Maximum Likelihood; Expectation Maximization

16. Unsupervised learning for clustering: other techniques [Murphy and Xu] {2 lectures}

    Similarity measures; evaluating clustering quality and choosing K; hierarchical and graph clustering (agglomerative, divisive, Bayesian)*

Other topics*

17. Optional selected topic(s) of student interest {~1 lecture}

e.g.:

1. Learn structure of data
2. Learn representation of the data for use in supervised learning.

* *As time permits.*

Key issues and concepts in ML

1. Hypothesis set (models being considered)
2. Objective function (fcn. being optimized)
3. Optimization method
4. Complexity (of model, data, and problem)
5. Assumptions and priors (inductive bias)

→ take in order.

1. Hypothesis set

In our house-price prediction example:
simplify → 1D input (x_i = living area)

Ex: Models: $\hat{f}_1(x) = w_0{}' + w_1{}'x$

$$\hat{f}_2(x) = w_0 + w_1 x + w_2 x^2$$

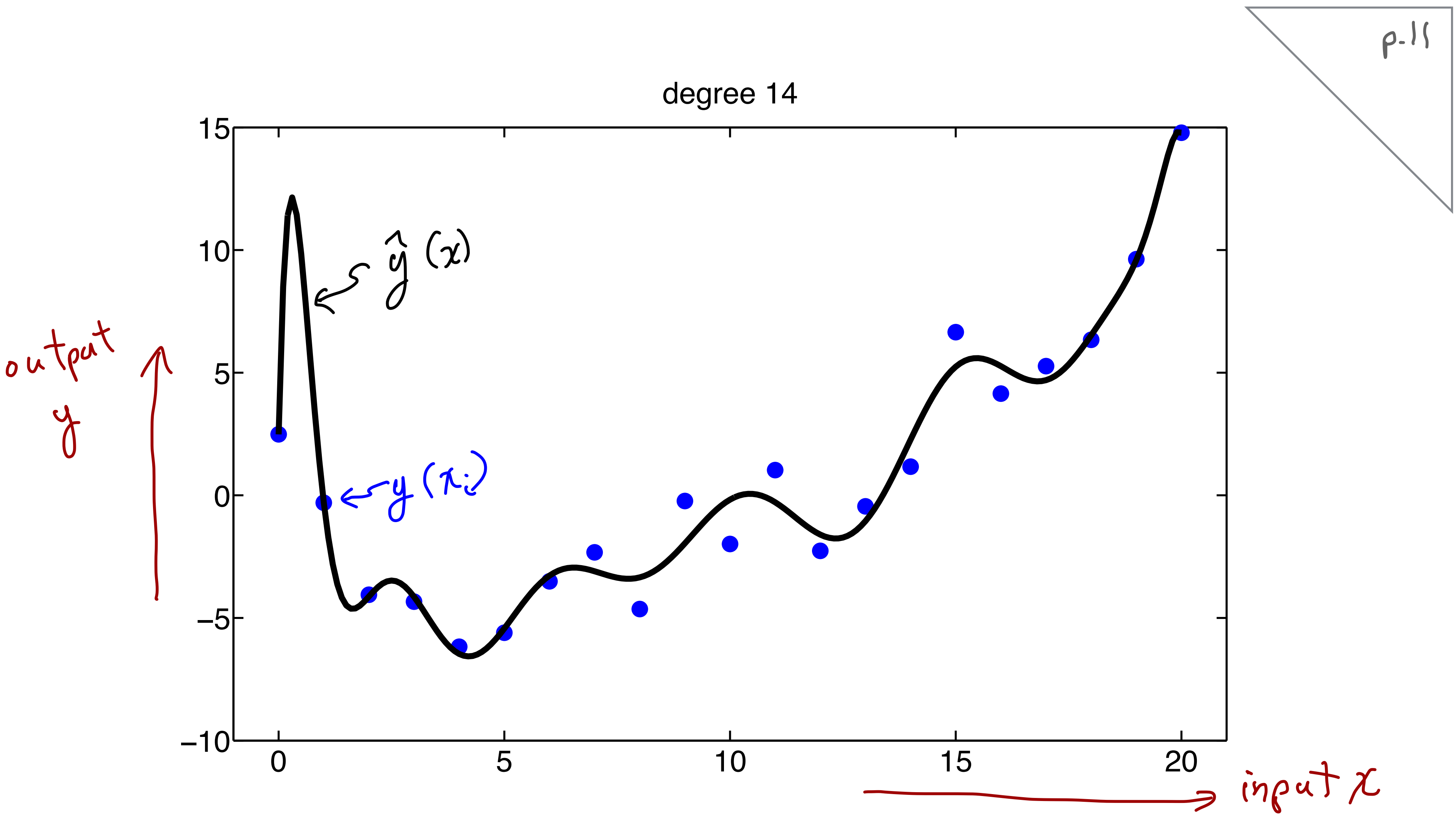
more generally: $\hat{f}_d(x) = \sum_{i=0}^{d} w_i x^i, \qquad 1 \le d \le d_{max}$

Our model selection & learning process chooses among all $d \ni 1 \le d \le d_{max}$, and all values of w_i.

⇒ Our hypothesis set, given $x \in \mathbb{R}$, $x \ge 0$:

$$\mathcal{H} = \left\{ \hat{f}_d(x) = \sum_{i=0}^{d} w_i x^i \;\middle|\; 1 \le d \le d_{max},\ d \in \mathbb{Z},\ w_i \in \mathbb{R} \right\}$$

[Fig.]



Murphy Fig. 1.18 (a). Regression to fit polynomial of degree 14, to 21 data points (minimizing MSE).

2. Objective function

→ Fcn. being optimized.

Ex: regression ex. (curve fit, house price, etc.):

$$J(\underline{w}, b) = MSE(\hat{y}_i, y_i) = \frac{1}{N_{Tr}} \sum_{i=1}^{N_{Tr}} (\hat{y}_i - y_i)^2$$

J is to be minimized.