Thursday, 9/10/2020

EE 660

MACHINE LEARNING FROM SIGNALS: FOUNDATIONS AND METHODS

Prof. B. Keith Jenkins

Lecture 6

p-1

Lecture 6	EE 660	Sep 10, 2020

Announcements

- Homework 1 is due Friday 9/11
- Homework 2 will be posted Friday 9/11

Today’s Lecture

- Notation: indicator function
- Logistic regression
  - Introduction
  - Logistic regression using MLE
    - Model
    - Objective function
    - Optimization
    - Regularization

Notation

Indicator function: $\mathbb{I}(\text{Expression}) = [\![\text{Expression}]\!] \triangleq \begin{cases} 1, & \text{if Expression} = \text{true} \\ 0, & \text{if Expression} = \text{false} \end{cases}$

Ex: $[\![g \geq 0]\!] = \begin{cases} 1 & \text{if } g \geq 0 \\ 0 & \text{if } g < 0 \end{cases}$

Sigmoid function: $\underbrace{\text{sigm}\{u\}}_{\text{Murphy}} = \underbrace{\theta(u)}_{\text{AML}} \triangleq \dfrac{e^u}{1+e^u}$

also called "logistic" function

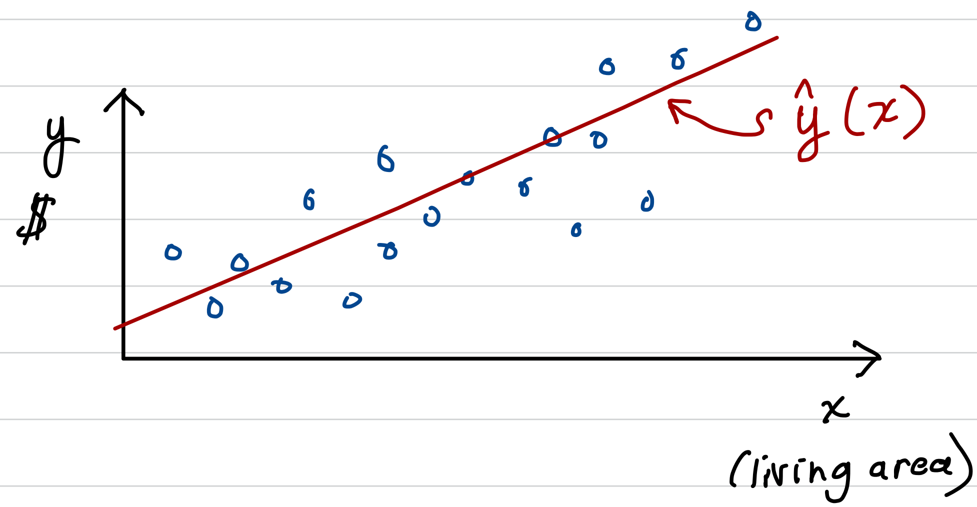
Logistic Regression [Murphy Ch.8] ($\underline{w} = \underline{w}^{(+)}$)

So far we have studied 2 realms of supervised learning:

- Regression
  Ex (1D input, linear model):

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

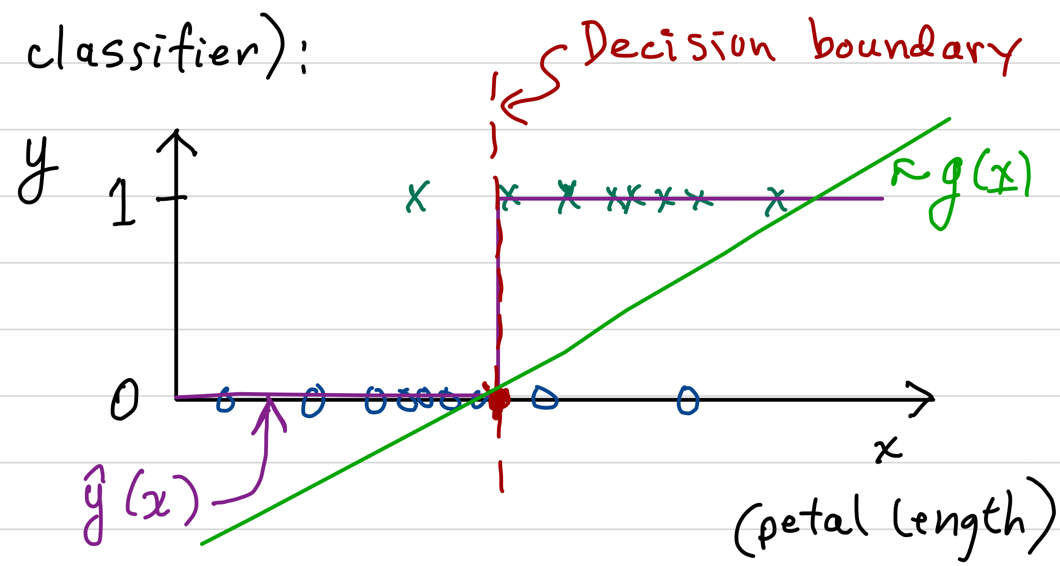
$$\left(\underline{x} = \begin{bmatrix} 1 \\ x \end{bmatrix} \right).$$



- Classification
  Ex (1D input, linear 2-class classifier):

$$\begin{cases} \text{o virginica } & y=0 \\ \text{x setosa } & y=1 \end{cases}$$

$$\hat{y}(x) = [\![g(x) \geq 0]\!]$$
$$= [\![\underline{w}^T \underline{x} \geq 0]\!]$$



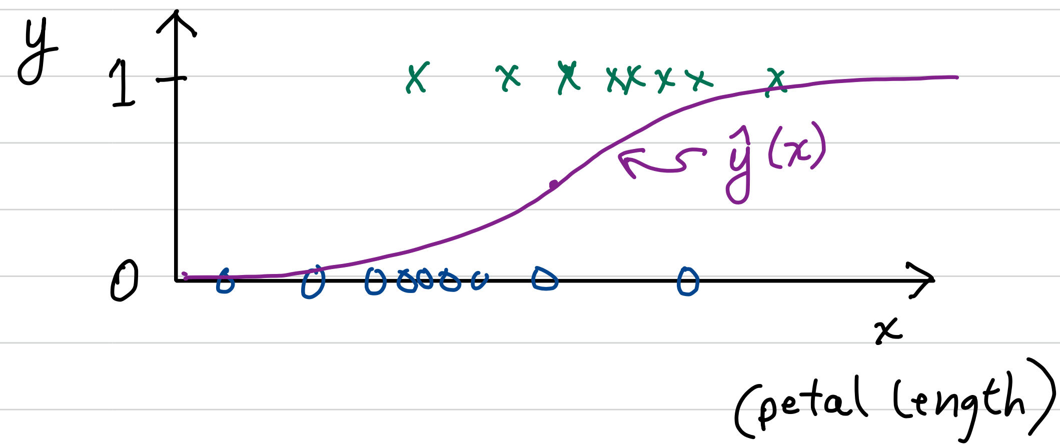
also plotted: $g(\underline{x}) = \underline{w}^T \underline{x} = w_0 + w_1 x$

Now consider:

- <u>Logistic Regression</u>

Ex.: same as classification above

$$\begin{cases} \text{o virginica} & y=0 \\ \text{x setosa} & y=1 \end{cases}$$



$$\hat{y}(x) = p(y=1 \mid x, \theta)$$
$$= p(\text{setosa} \mid x, \theta)$$
$$= \text{sigm}\{\underline{w}^T \underline{x}\}$$

Comments:

1. $\hat{y}(x)$ is <u>not</u> trying to mimic or approximate the data.

2. Logistic regression is a form of <u>classification</u>.

Logistic Regression for Supervised ML

1. Model [M 8.2]

$$p(y|\underline{x},\underline{w}) = \mathrm{Ber}(y|\,\mathrm{sigm}\{\underline{w}^T\underline{x}\})$$

$$= \mu^{\mathbb{I}(y=1)}(1-\mu)^{\mathbb{I}(y=0)}$$

in which $\mu = \mathrm{sigm}\{\underline{w}^T\underline{x}\}$

Change output representation

$y \in \{0, 1\}$

Let $\tilde{y} = 2y - 1 \Rightarrow \tilde{y} \in \{-1, +1\}$

$$p(\tilde{y}|\underline{x},\underline{w}) = \mu^{\mathbb{I}(\tilde{y}=1)}(1-\mu)^{\mathbb{I}(\tilde{y}=-1)}$$

$$p(\tilde{y}_i|\underline{x}_i,\underline{w}) = [\mathrm{sigm}(\underline{w}^T\underline{x}_i)]^{\mathbb{I}(\tilde{y}_i=1)}\,[\underbrace{1-\mathrm{sigm}(\underline{w}^T\underline{x}_i)}_{1-\mathrm{sigm}(-\tilde{y}_i\underline{w}^T\underline{x}_i)}]^{\mathbb{I}(\tilde{y}_i=-1)}$$

Use: $\mathrm{sigm}(s) = 1-\mathrm{sigm}(-s)$

$\Rightarrow \mathrm{sigm}(\tilde{y}_i\underline{w}^T\underline{x}_i)$

$$p(\tilde{y}_i|\underline{x}_i,\underline{w}) = [\mathrm{sigm}(\tilde{y}_i\underline{w}^T\underline{x}_i)]^{\mathbb{I}(\tilde{y}_i=1)}\,[\mathrm{sigm}(\tilde{y}_i\underline{w}^T\underline{x}_i)]^{\mathbb{I}(\tilde{y}_i=-1)}$$

$$\boxed{\Rightarrow \quad p(\tilde{y}_i \mid \underline{x}_i, \underline{w}) = \mathrm{sigm}(\tilde{y}_i \underline{w}^T \underline{x}_i)}$$

2. Objective function

→ Use maximum likelihood

Likelihood: $\quad p(\mathcal{D} \mid \underline{w}) = p(\underline{\tilde{y}} \mid \underline{\underline{X}}, \underline{w}) = \prod_{i=1}^{N} p(\tilde{y}_i \mid \underline{x}_i, \underline{w})$

$$= \prod_{i=1}^{N} \frac{e^{\tilde{y}_i \underline{w}^T \underline{x}_i}}{1 + e^{\tilde{y}_i \underline{w}^T \underline{x}_i}} \cdot \left(\frac{e^{-(-)}}{e^{-(--)}} \right)$$

$$= \prod_{i=1}^{N} \frac{1}{e^{-\tilde{y}_i \underline{w}^T \underline{x}_i} + 1}$$

$$\boxed{-\ell(\underline{w}) = NLL(\underline{w}) = \sum_{i=1}^{N} \underbrace{\ln\left[1 + e^{-\tilde{y}_i \underline{w}^T \underline{x}_i}\right]}_{E_i} = J(\underline{w}, \mathcal{D})}$$

objective fcn. for MLE of $\underline{w}$ in logistic regression.

$J(\underline{w}, \mathcal{D})$ is differentiable

$J(\underline{w}, \mathcal{D})$ is convex

Interpretation: $J(\underline{w}, \mathcal{D}) = \sum_{i=1}^{N} \underbrace{E_i}$

Note that:

$\tilde{y}_i \underline{w}^T \underline{x}_i > 0 \Rightarrow$ correct classification

$< 0 \Rightarrow$ incorrect "

E_i acts acts as a (continously varying) error or loss term for the i^{th} data point, given $\underline{w}$.

3. Optimization

Can we $\nabla_{\underline{w}} J(\underline{w}, \mathcal{D}) = \underline{0}$ & solve algebraically?

→ not amenable to this approach.

Use gradient-based techniques:

SGD

Batch GD

Mini-batch GD

Newton's method

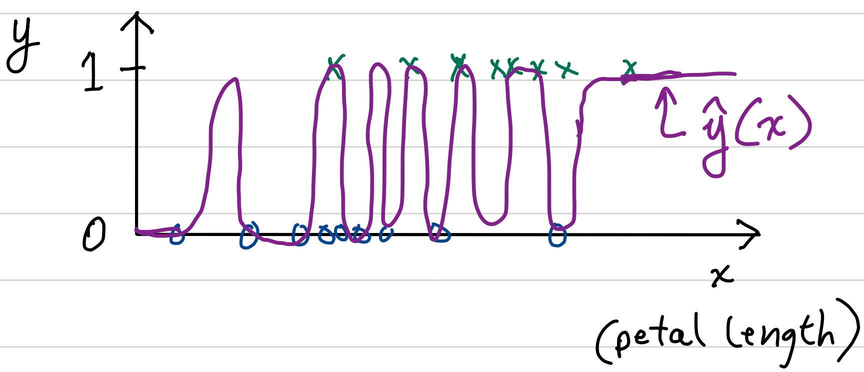
Iterative Reweighted Least Squares (IRLS) [M 8.3.4]

(N.R.F.)

4. Complexity; assumptions & priors.

→ Can logistic regression overfit? Yes.

Ex:



Use $\mu = \text{sigm}\{\underline{w}^T \underline{\phi}(\underline{x})\}$

$\underline{\phi}(\underline{x})$ = nonlinear fcn. of $\underline{x}$.

→ Can we use a regularizer? Yes.

$$\tilde{J}(\underline{w}, \mathcal{D}) = \underbrace{NLL(\underline{w})}_{J(\underline{w}, \mathcal{D})} + \underbrace{\lambda \|\underline{w}\|_2^2}_{\text{also convex } (\lambda \geq 0).}$$