Announcements

- Homework 3 has been posted.

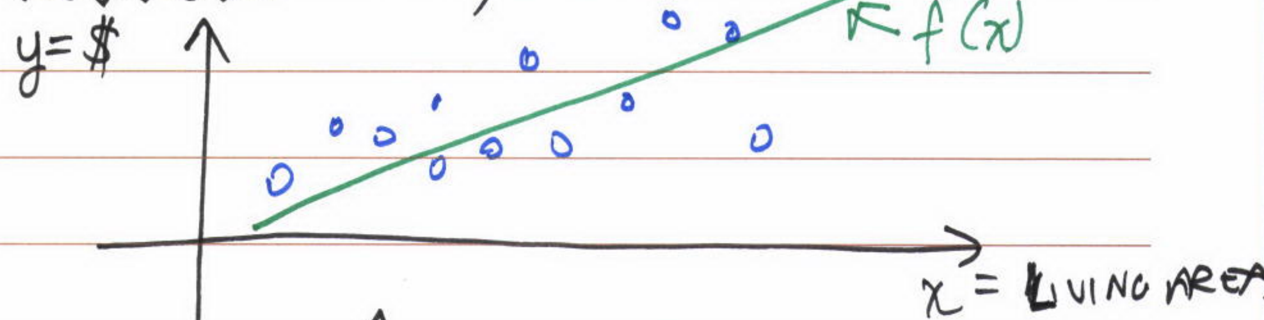
Today's Lecture

- Logistic regression

LOGISTIC REGRESSION

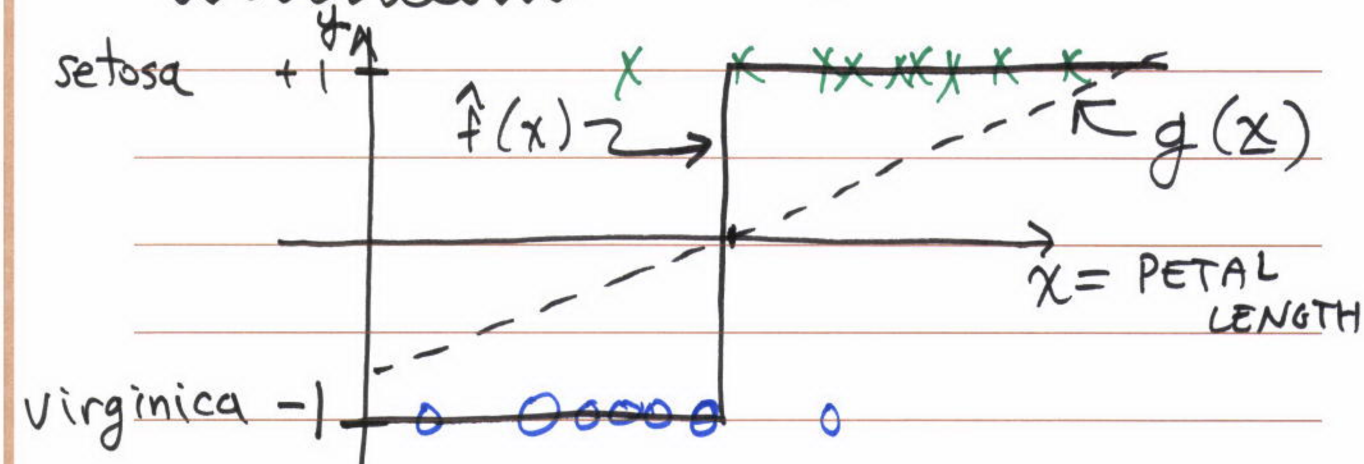
INTRO. (AUGMENTED NOTATION)

REGRESSION (1D, LINEAR)



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

CLASSIFICATION (LINEAR)



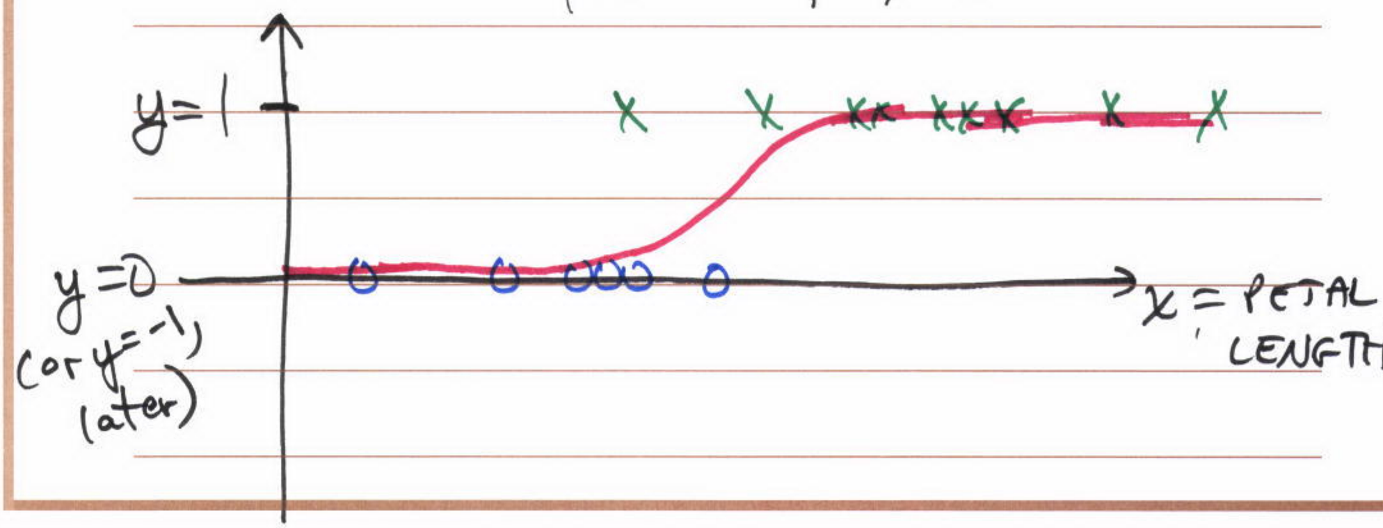
$$g(\underline{x}) = \underline{w}^T \underline{x}$$

$$\hat{y}(\underline{x}) = \hat{f}(\underline{x}) = \operatorname{sgn}\{g(\underline{x})\} = \operatorname{sgn}\{\underline{w}^T \underline{x}\}$$

Logistic Regression

$$\hat{f}(x) = p(y=1 \mid x, \mathcal{D})$$

$$= p(\text{setosa} \mid x, \mathcal{D})$$



$$\hat{f}(\underline{x}) = \text{sigm}\{\underline{w}^T \underline{x}\}$$ [Murphy notation]

$$= \frac{e^{\underline{w}^T \underline{x}}}{1 + e^{\underline{w}^T \underline{x}}}$$

$$= \theta(\underline{w}^T \underline{x})$$ [AML notation]

= "logistic" or "sigmoid" function (of $\underline{w}^T \underline{x}$)

MODEL

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

$$= \mu^{\mathbb{I}(y=1)} (1-\mu)^{\mathbb{I}(y=0)} \qquad [\text{M } 8.2]$$

IN WHICH $\mu = \mathrm{sigm}(\underline{w}^T \underline{x})$

CHANGE OUTPUT (y) REPRESENTATION:

LET $\tilde{y} = 2y - 1 \implies \tilde{y} \in \{-1, +1\}$

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

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

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

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

OBJECTIVE FCN.

MAX. LIKELIHOOD

$$\text{LIKELIHOOD} = p(\mathcal{D} \mid \underline{w})$$

$$= \prod_{i=1}^{N} p(\tilde{y}_i \mid \underline{x}_i, \underline{w})$$

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

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

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

↑ OBJ. FCN. FOR MLE OF $\underline{w}$ IN LOGISTIC REGRESSION.

J IS CONVEX.

LET: $-\ell(\underline{w}) = \sum_{i=1}^{N} E_i$, $\quad E_i = \ln\left[1 + e^{-\tilde{y}_i \underline{w}^T \underline{x}}\right]$

$\tilde{y}_i$	$\underline{w}^T \underline{x}_i$	$E_i = \ln\left[1 + e^{-\tilde{y}_i \underline{w}^T \underline{x}}\right]$
$+1$	>0	$0 < E_i < \ln 2$
$+1$	$\gg 0$	$E_i \gtrsim 0$
$+1$	<0	$E_i > \ln 2$
$+1$	$\ll 0$	$E_i \gg \ln 2$
$\vdots$		$\hookrightarrow E_i \approx - \underline{w}^T \underline{x}$

<u>HOW TO MINIMIZE J</u> ? $\rightarrow J$ IS CONVEX.

CAN WE $\nabla_{\underline{w}} J(\underline{w}, \mathcal{D}) = 0$, SOLVE ALGEBRAICALLY?

$\rightarrow$ NOT TRACTIBLE.

HOW ELSE?

- ITERATIVE, GRADIENT-BASED TECHNIQUES.
  - > GRADIENT DESCENT (STEEPEST DESCENT)

$$\underline{w}(k+1) = \underline{w}(k) - \eta(k) \nabla_{\underline{w}} J$$

(BATCH UPDATE)

> STOCHASTIC GRADIENT DESCENT

(i) INITIALIZING $\underline{w}(0)$

(ii) ITERATE OVER DATA POINTS UNTIL CONVERGENCE:

(a) RANDOMLY PICK A DATA POINT (WITH REPLACEMENT)

(b) PERFORM A SINGLE-SAMPLE UPDATE:

$$\underline{w}(i+1) = \underline{w}(i) - \eta(i) \nabla_{\underline{w}} E_i$$

(i = ITERATION NUMBER)

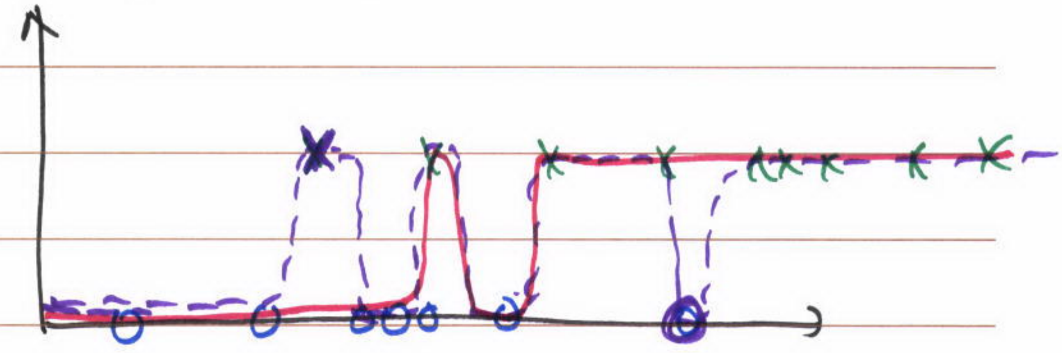
CAN SHOW:

$$\nabla_{\underline{w}} E_i = \frac{\tilde{y}_i \underline{x}_i}{1 + e^{\tilde{y}_i \underline{w}^T(i) \underline{x}_i}}$$

(e.g., see EE559 or Discussion 3).

$\exists$ OTHER TECHNIQUES (e.g., Murphy 8.3.2–8.3.5).

- IS REGULARIZATION USEFUL (AND POSSIBLE) IN LOGISTIC REGRESSION?
  - CAN L.R. OVERFIT?



$$\hat{f}(x) = \mathrm{sigm}\left\{ \underline{w}^T \underline{\phi}(\underline{x}) \right\}$$

↑ POLYNIAL IN x TERMS

→ YES, WOULD BE USEFUL.

→ YES, WE CAN ADD A PRIOR TERM $p(\underline{w})$,

⇒ MAP EST. [⟺ RIDGE REGRESSION]

OR BAYESIAN INFERENCE.

→ LOGISTIC REGRESSION:

$$\rightarrow J(\underline{w}, \mathcal{D}) = \underbrace{NLL(\underline{w})}_{\text{from logistic regr., MLE.}} + \lambda \|\underline{w}\|_2^2$$

(IF $p(\underline{w})$ = GAUSSIAN $= N(\underline{w} \mid 0, \tau^2)$)