Tuesday, 9/8/2020

EE 660

MACHINE LEARNING FROM SIGNALS: FOUNDATIONS AND METHODS

Prof. B. Keith Jenkins

Lecture 5

p. 1

Lecture 5 | **EE 660** | **Sep 8, 2020**

Announcements

- Homework 2 was posted; due this Friday at 5:00 PM PDT.
- Supplemental Video 1 (Week 2) on Ridge Regression was posted last Friday; its content is required for this class

Today's Lecture

- Notation comment: data vs. variables
- Discriminative and generative models
- Bayesian inference
  - For discriminative models
  - For generative models

Notation Comment – Data vs. variables

General variables

$\underline{x}$ = input (feature vector), y = output (value or class)

x_j = $j^{\underline{th}}$ component of feature vector $\underline{x}$

Which are different than:

Data points (of dataset)

$\underline{x}_i$ = input (feature) values of $i^{\underline{th}}$ data point of $\mathcal{D}$

y_i = output (value or class) of $i^{\underline{th}}$ data point of $\mathcal{D}$

$$\underline{\underline{X}} = \left\{ \begin{array}{l} \text{input values} \\ \text{of all points} \\ \text{in data set } \mathcal{D} \end{array} \right\} = \begin{bmatrix} \underline{x}_1^T \\ \underline{x}_2^T \\ \vdots \\ \underline{x}_N^T \end{bmatrix}, \quad \underline{y} = \left\{ \begin{array}{l} \text{output values} \\ \text{of all points} \\ \text{in dataset } \mathcal{D} \end{array} \right\} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

$N \times D$ or
$N \times (D+1)$

Discriminative and Generative Models [1]

Generative approach models $p(\underline{x}, y)$ or $p(\underline{x}, y \mid \underline{\theta})$.

Then $p(y \mid \underline{x})$ or $p(y \mid \underline{x}, \underline{\theta})$ can be obtained, e.g.:

$$p(y \mid \underline{x}) = \frac{p(\underline{x}, y)}{p(\underline{x})} \quad \text{for } p(\underline{x}) \neq 0.$$

→ Common in classification

Discriminative approach models $p(y \mid \underline{x})$ or $p(y \mid \underline{x}, \underline{\theta})$ directly.

→ Common in regression and classification

What if we model $p(\underline{x} \mid y)$ or $p(\underline{x} \mid y, \underline{\theta})$?

→ If we also have $p(y)$, then $p(\underline{x}, y) = p(\underline{x} \mid y)\, p(y)$

⇒ It's generative

[1] For more information, see Murphy 8.6.0 – 8.6.1

Bayesian Inference ($\underline{w} = \underline{w}^{(t)}$) [Murphy 7.6]

(Murphy uses both $\underline{w}^{(0)}$ and $\underline{w}^{(t)}$ in 7.6)

We want to estimate $\underline{\theta}$.

→ Instead of finding a point estimate $\hat{\underline{\theta}}$, we will estimate the <u>density</u>:

$$p(\underline{\theta} \mid \mathcal{D}) = \text{"parameter posterior"}$$

How?

What did we do previously, to find $\hat{\underline{\theta}}_{MLE}$?

For discriminative approach:

Model $\quad p(y \mid \underline{x}, \underline{\theta}) \qquad$ [Ex: $p(y \mid \underline{x}, \underline{\theta}) = \mathcal{N}(y \mid \underline{w}^T \underline{x}, \sigma^2)$]

Use training data $\mathcal{D}$ to get log likelihood

$$\ln p(\mathcal{D} \mid \underline{\theta}) = \sum_{i=1}^{N} \ln p(y_i \mid \underline{x}_i, \underline{\theta})$$

Optimize : $\hat{\underline{\theta}}_{MLE} = \underset{\underline{\theta}}{\operatorname{argmin}} \left\{ -\sum_{i=1}^{N} \ln p(y_i \mid \underline{x}_i, \underline{\theta}) \right\}$

How to find $p(\underline{\theta} \mid \mathcal{D})$?

We use a model for: $p(y \mid \underline{x}, \underline{\theta})$ (discriminative approach)
or $p(\underline{x} \mid y, \underline{\theta})$ (generative approach

Ex: $p(y \mid \underline{x}, \underline{\theta}) = N(y \mid \underline{w}^T \underline{x}, \sigma^2)$, $\underline{\theta} = \underline{w}$ or $\underline{\theta} = \begin{bmatrix} \underline{w} \\ \sigma^2 \end{bmatrix}$.

(linear, Gaussian model for regression)

Use training dataset $\mathcal{D}$ with the model, to get likelihood: $p(\mathcal{D} / \underline{\theta})$

Then use Bayes' theorem:

$$(1) \quad \boxed{p(\underline{\theta} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \underline{\theta})\, p(\underline{\theta})}{p(\mathcal{D})}, \quad p(\mathcal{D}) = \int p(\mathcal{D} \mid \underline{\theta}')\, p(\underline{\theta}')\, d\underline{\theta}'}$$

(or sum if $\underline{\theta}'$ is discrete)

Ex: Housing price prediction

$$p(\underline{\theta} \mid \mathcal{D}) = p(\underline{w} \mid \mathcal{D}), \qquad \underline{w} = \begin{bmatrix} \text{Living area} \\ \text{\# of rooms} \\ \text{Age of house} \\ \vdots \end{bmatrix}$$

$$p(\mathcal{D} \mid \underline{\theta}) \propto \prod_{i=1}^{N} p(y_i \mid \underline{x}_i, \underline{\theta}) = \underbrace{\prod_{i=1}^{N} N(y_i \mid \underline{w}^T \underline{x}_i, \sigma^2)}_{\text{known fcn. of } \underline{w}}$$

$$p(\underline{\theta}) = p(\underline{w}) = \text{prior on } \underline{w} \quad \leftarrow \text{we will specify}$$

Plug into (1) $\Rightarrow p(\underline{\theta} \mid \mathcal{D}) = p(\underline{w} \mid \mathcal{D})$

To make predictions, we want $p(\,?\,|\,?\,)$ $\rightarrow p(y|\underline{x}, \mathcal{D})$

$= \underline{\text{Posterior predictive}}$

From: $p(y) = \int p(y|\underline{\theta})\, p(\underline{\theta})\, d\underline{\theta}$

We get:

$$p(y|\underline{x}, \mathcal{D}) = \int p(y|\underline{x}, \underline{\theta}, \mathcal{D})\, p(\underline{\theta}|\underline{x}, \mathcal{D})\, d\underline{\theta}$$

If we're given $\underline{\theta}$ and $\underline{x}$, then $\mathcal{D}$ provides no useful information to predict y.

If we're given $\mathcal{D}$, then $\underline{x}$ provides no useful information for $\underline{\theta}$.

(2) $\therefore$ $$\boxed{p(y|\underline{x}, \mathcal{D}) = \int p(y|\underline{x}, \underline{\theta})\, p(\underline{\theta}|\mathcal{D})\, d\underline{\theta}}$$

$$\underbrace{p(y|\underline{x}, \underline{\theta})}_{\text{from our model (see below)}} \quad \underbrace{p(\underline{\theta}|\mathcal{D})}_{\text{from (1).}}$$

Comments

1. Eq. (2) is a weighted average of the density of y for each value of $\underline{\theta}$, weighted by the density of $\underline{\theta}$ given $\mathcal{D}$.

2. If instead we used the peak value of $p(\underline{\theta} | \mathcal{D})$ as in:

$$\hat{\underline{\theta}}_{MAP} = \underset{\underline{\theta}}{\operatorname{argmax}} \left\{ p(\underline{\theta} | \mathcal{D}) \right\}$$

then our posterior predictive would be just:

$$p(y | \underline{x}, \mathcal{D}) = p(y | \underline{x}, \underline{\theta} = \hat{\underline{\theta}}_{MAP})$$

Getting $p(y | \underline{x}, \underline{\theta})$ from our model

(a) Discriminative case : e.g. $p(y | \underline{x}, \underline{\theta}) = \mathcal{N}(y | \underline{w}^T \underline{x}, \sigma^2)$

(b) Generative case: let original model specify $p(\underline{x} | y, \underline{\theta})$

Then:

$$p(y | \underline{x}, \underline{\theta}) = \frac{p(\underline{x} | y, \underline{\theta})\, p(y)}{p(\underline{x})}$$

e.g.: $p(\underline{x} | y, \underline{\theta}) = \mathcal{N}(\underline{x} | y, \underline{m}_y, \underline{\underline{\Sigma}}_y)$ [EE559]

$p(y_i | x_i, \underline{w})$

likelihood

Based on i^{th} data point only.

Parameter
prior/posterior

data space

prior $p(\underline{w})$ = Gaussian

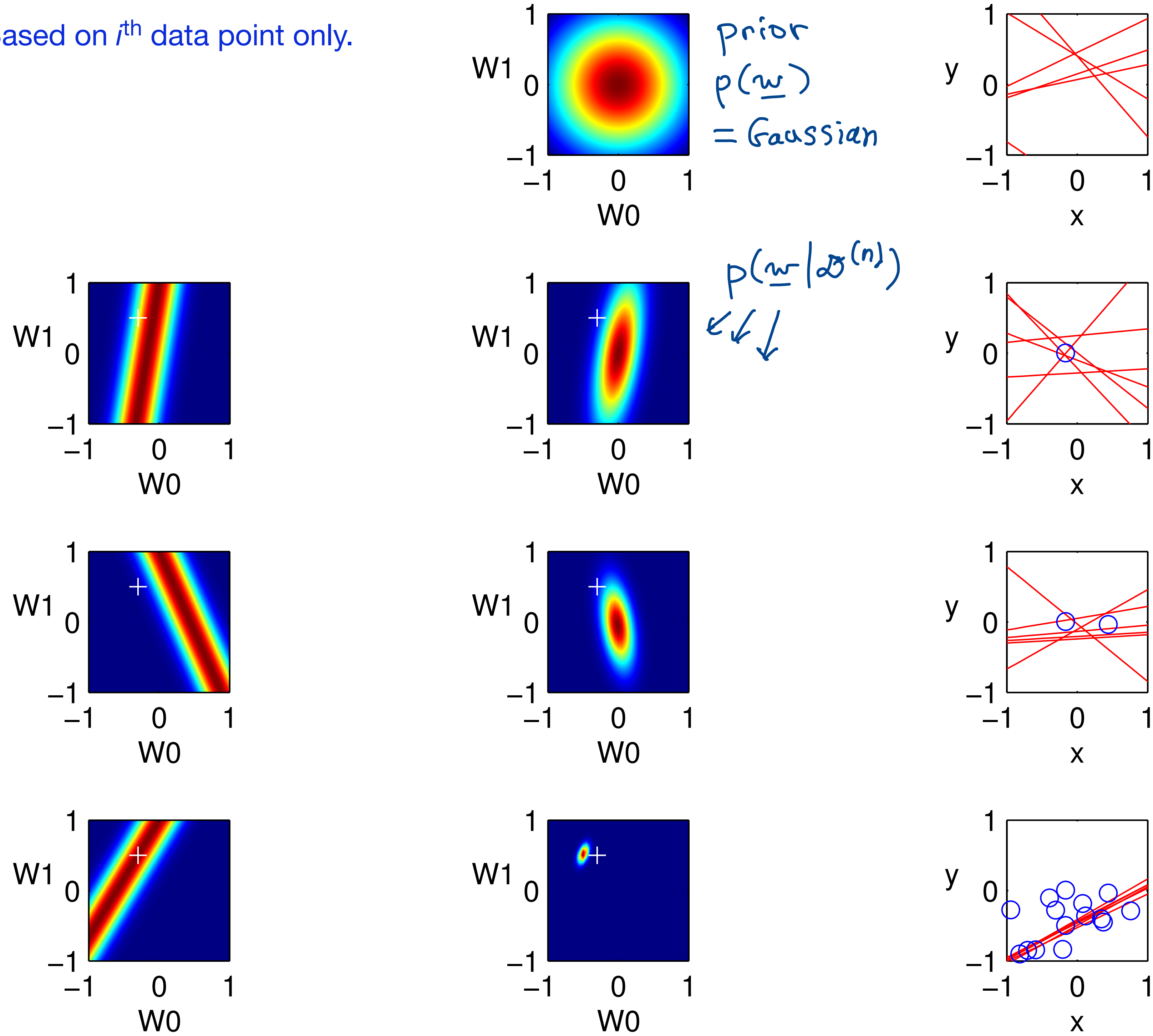
$n=0$

$p(\underline{w} | \mathcal{D}^{(n)})$

n=1

n=2

n=20



○ data point

Lines drawn from samples of posterior predictive $p(y | x, \mathcal{D}^{(n)})$. Each sample $\Rightarrow y_s(x)$.

Murphy Fig. 7.11.

$\uparrow$

$\ln p(y_i | x_i, \underline{w})$

$\propto -\left[y_i - (w_0 + w_1 x_i)\right]^2 = 0$

(at peak)

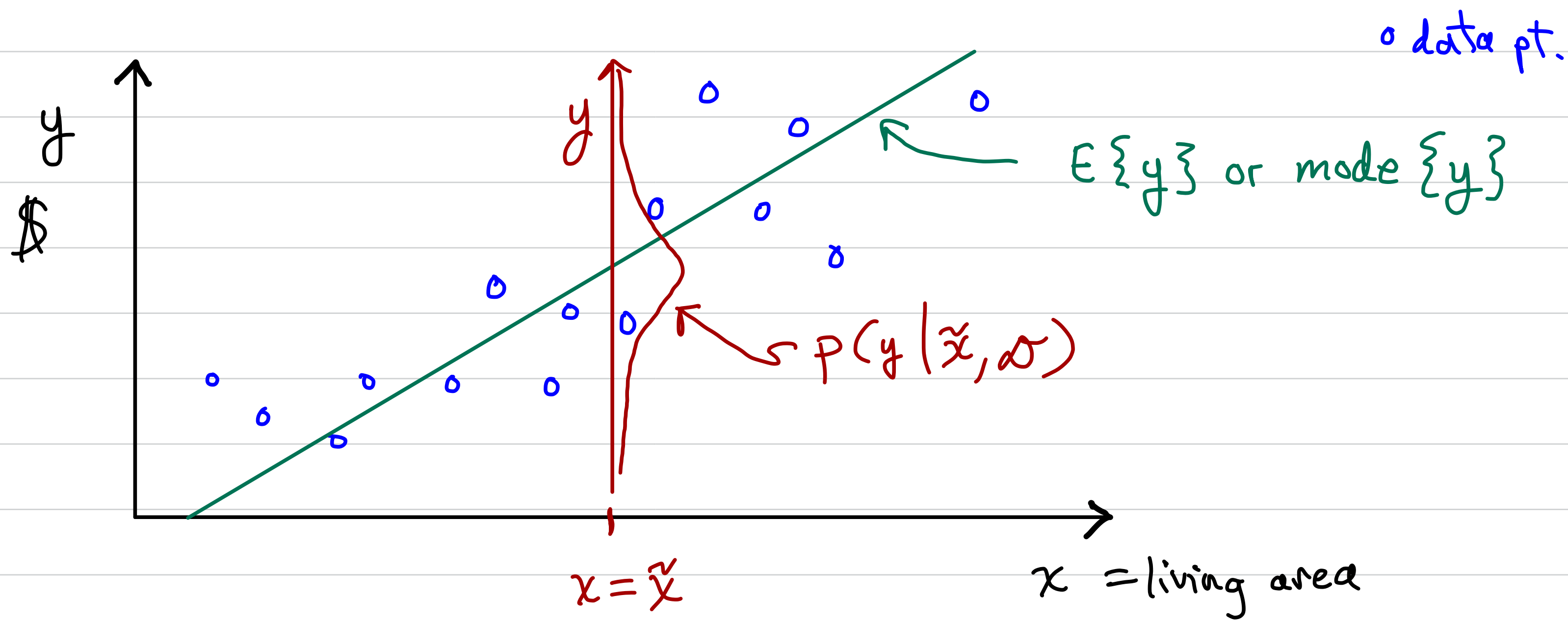
$\Rightarrow y_i = (w_0 + w_1 x_i)$ is a line in $\underline{w}$-space.

Model is discriminative, Gaussian linear regression:

$p(y | x, \underline{w}) = \mathcal{N}(y | (w_0 + w_1 x), \sigma^2)$

Interpreting final prediction (posterior predictive)

$p(y \mid \underline{x}, \mathcal{D})$. Let $y = \$$ house, $\underline{x} = x =$ living area



= living area of house with unknown price.