Gaussians and Linear Regression

Machine Learning - 10601

Geoff Gordon, Miroslav Dudík

http://www.cs.cmu.edu/~ggordon/10601/

September 16, 2009

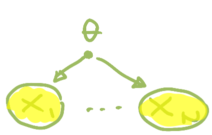
Last time: density estimation

POINT ESTIMATION

Input: $X_1, X_2, \ldots, X_N$ i.i.d.

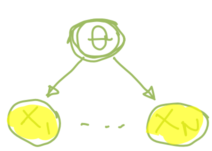
Output: p(X)

Maximum likelihood: $\max_{\theta}\ p(\text{data} \mid \theta) = \prod_n p(x_n \mid \theta)$ ($x_1, \ldots, x_n$)



$\hat{\theta}_{ML} \leftarrow \max_{\theta} \underbrace{\sum_n \log p(x_n \mid \theta)}_{\text{GOODNESS OF FIT}}$

Maximum a posteriori (MAP): $\max_{\theta}\ p(\theta \mid \text{data}) = \frac{p(\theta) . p(\text{data} \mid \theta)}{\cancel{p(\text{data})}}$ const.



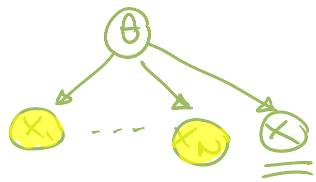
$\hat{\theta}_{MAP} \leftarrow \max_{\theta} \underbrace{\sum_n \log p(x_n \mid \theta)}_{\text{GOODNESS OF FIT}} + \underbrace{\log p(\theta)}_{\text{LEVEL OF BELIEF}}$

Last time: density estimation

Input: $X_1, X_2, \ldots, X_N$

Output: $p(X)$

Bayesian prediction:



$$p(x \mid \overbrace{x_1, \ldots, x_n}^{\text{data}})$$

$$\propto \int p(x \mid \theta)\, p(\theta \mid \text{data})\, d\theta$$

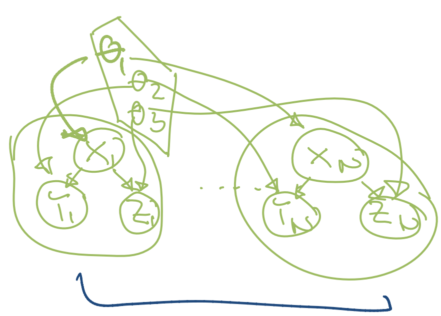
IN GENERAL: DIFFICULT

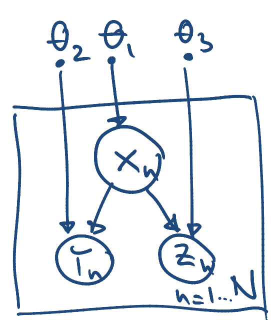
CONJUGATE PRIORS

BINOM & BETA

Density estimation

- we used the thumbtack example
- similar for structured variables
  - diseases and symptoms; natural images; natural language…





Continuous variables: Gaussian

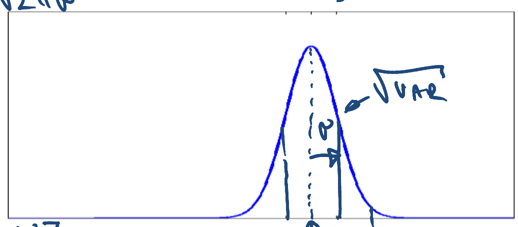
SALARY OF A FRESH CMU GRADUATE

$$\mathcal{N}(X \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

$x \in \mathbb{R}$

MEAN $= E[x]$

VARIANCE $= E[(x-\mu)^2]$



Obtaining Gaussians from Gaussians

NORMAL DIST.

- **adding constants and scaling**

  $X \sim \mathcal{N}(X \mid \mu, \sigma^2)$

  $Y = aX + b$

  $Y \sim \mathcal{N}(Y \mid a\mu + b,\ a^2\sigma^2)$

- **summing independent Gaussians**

  $X \sim \mathcal{N}(X \mid \mu_X, \sigma_X^2)$

  $Y \sim \mathcal{N}(Y \mid \mu_Y, \sigma_Y^2)$

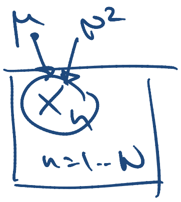
  $Z = X + Y$

  $Z \sim \mathcal{N}(Z \mid \mu_X + \mu_Y,\ \sigma_X^2 + \sigma_Y^2)$

  $E[Z] = \underbrace{E[X]}_{=\mu_X} + \underbrace{E[Y]}_{=\mu_Y}$

Learning Gaussians

Maximum Likelihood



$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

$$\max_{\mu,\sigma^2} \sum_n \log p(x_n|\mu,\sigma^2)$$

$$\sum_n \left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x_n-\mu)^2}{2\sigma^2}\right]$$

$$\frac{\partial}{\partial\mu} = \sum_n \left[+\frac{2(x_n-\mu)}{2\sigma^2}\right] = 0$$

$$E[\hat{\mu}_{ML}] = \mu$$

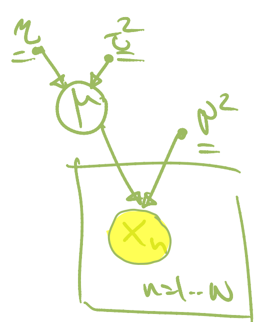
$$\sum_n x_n = N\mu$$

$$\hat{\mu}_{ML} = \bar{x} = \frac{1}{N}\sum_n x_n$$

BIASED ESTIMATOR

$$\frac{\partial}{\partial\sigma} = \rightsquigarrow \hat{\sigma}^2_{ML} = \frac{1}{N}\sum_n (x_n - \hat{\mu}_{ML})^2 \qquad E[\hat{\sigma}^2_{ML}] = \frac{N-1}{N}\sigma^2$$

Posterior of the mean



$$x \sim \mathcal{N}(x|\mu,\sigma^2)$$

KNOWN

$$\mu \sim \mathcal{N}(\mu|\eta,\tau^2)$$

$$p(\mu|x_1,\dots,x_n,\eta,\tau^2,\sigma^2)$$

$$= \frac{p(\mu|\eta,\tau^2)\,p(x_1,\dots,x_n|\mu,\sigma^2)}{\underbrace{p(x_1,\dots,x_n|\eta,\tau^2,\sigma^2)}_{\text{CONST.}}}$$

$$p(\mu,x_1,\dots,x_n|\eta,\tau^2,\sigma^2)$$

$$\propto \exp\left\{-\frac{(\mu-\eta)^2}{2\tau^2}\right\}\prod_n \exp\left\{-\frac{(x_n-\mu)^2}{2\sigma^2}\right\}$$

$$= \exp\left\{-\mu^2\cdot(\dots) + \mu\cdot(\dots) + \dots\right\} = \mathcal{N}(\mu|\hat{\eta},\hat{\tau}^2)$$

Posterior of the mean

POSTERIOR: $\mathcal{N}(\mu \mid \tilde{\mu}, \tilde{\tau}^2)$

PRIOR

SAMPLE AVG

$$\tilde{\mu} = \frac{\frac{1}{\tau^2}\mu + \frac{N}{\sigma^2}\bar{x}}{\frac{1}{\tau^2} + \frac{N}{\sigma^2}}$$

$\tau^2 \approx 0$

$\tau^2 \approx \infty$

WEIGHTED HARMONIC AVG of $\underbrace{\tau^2 \ \& \ \sigma^2}$

$$\tilde{\tau}^2 = \frac{1}{1+N} \cdot \frac{1+N}{\frac{1}{\tau^2} + \frac{N}{\sigma^2}}$$

POSTERIOR VARIANCE DECREASES w/ SAMPLE SIZE

Things to know about Gaussians

- $\hat{\mu}_{MLE} = \frac{1}{N}\sum_n x_n$

  $\hat{\sigma}^2_{MLE} = \frac{1}{N}\sum_n (x_n - \hat{\mu}_{MLE})^2 \quad \| \quad \underbrace{\frac{1}{N-1}\sum_n \ldots}$ $E[\ldots] = \sigma^2$
- conjugate for the mean: Gaussian
  - variance of posterior decreases with number of samples
  - mean of posterior gets closer and closer to sample average

Gaussians

- modeling singleton continuous variables

NEXT:

- modeling dependence among continuous variables:
  - stock prices given previous few days
  - salary given GPA

  …

NOT IN X; IN W

Linear regression

$t = w_0 + w_1 \cdot GPA + w_2 \cdot GPA^2$

GPA SALARY

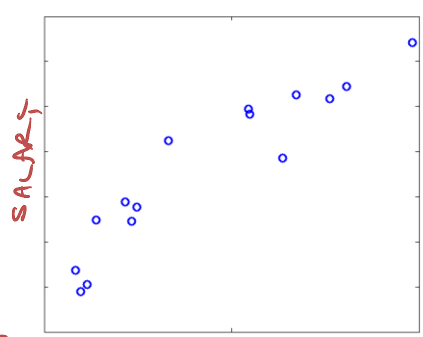
Input: $(\mathbf{x}_1,\mathbf{t}_1)$, $(\mathbf{x}_2, \mathbf{t}_2)$… , $(\mathbf{x}_N, \mathbf{t}_N)$

Goal: learn a map $\mathbf{x} \mapsto \mathbf{t}$

SALARY

$\mathbf{t} \approx \mathbf{w}_0 + \mathbf{w}_1\mathbf{x} + \mathbf{w}_2\mathbf{x}^2$

$\boxed{\text{LINEAR}}$ IN w_0, w_1, w_2



$t \approx \sum_j w_j \phi_j(x)$ ← BASIS FCTS, FEATURES

$\phi_0(x) = 1$

$\phi_1(x) = GPA$

$\phi_2(x) = GPA^2$

$\phi_3(x) = GENDER \begin{cases} 0 \text{-MALES} \\ 1 \text{-FEMALES} \end{cases}$

$\phi_4(x) = GENDER \cdot GPA$

Linear regression

Input: $(x_1,t_1), \dots, (x_N, t_N)$

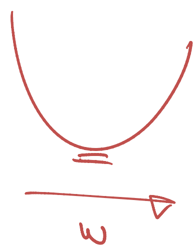
Assume: $t \approx \sum_j w_j \phi_j(x)$

Goal: $\min_{w} \sum_n (t_n - \underline{w} \cdot \underline{\phi}(x_n))^2$

INNER PROD

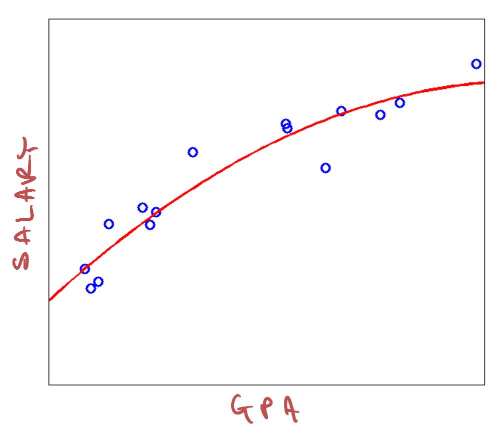
$\nabla_w = 0$

$\hat{w}_{LSE}$



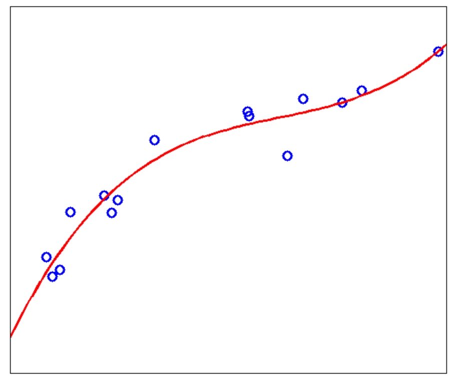
Linear regression

$t \approx \underline{w_0} + \underline{w_1} x + \underline{w_2} x^2$



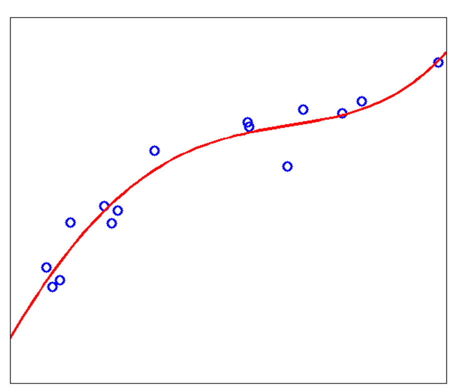
Linear regression

$t \approx w_0 + w_1x + w_2x^2$
$+ w_3x^3$



Linear regression

$t \approx w_0 + w_1x + w_2x^2$
$+ w_3x^3 + w_4x^4$

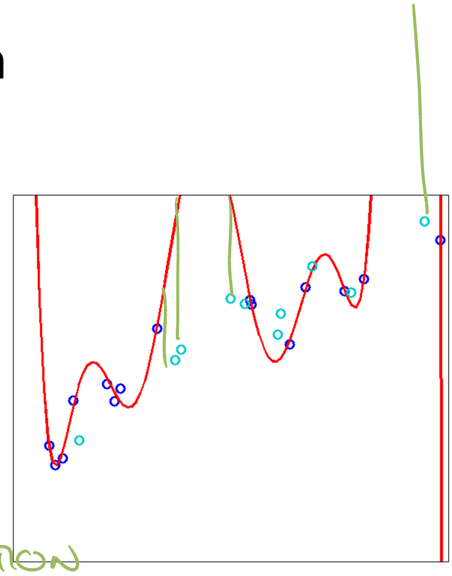


Linear regression

$$t \approx w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + \dots + w_{10} x^{10}$$

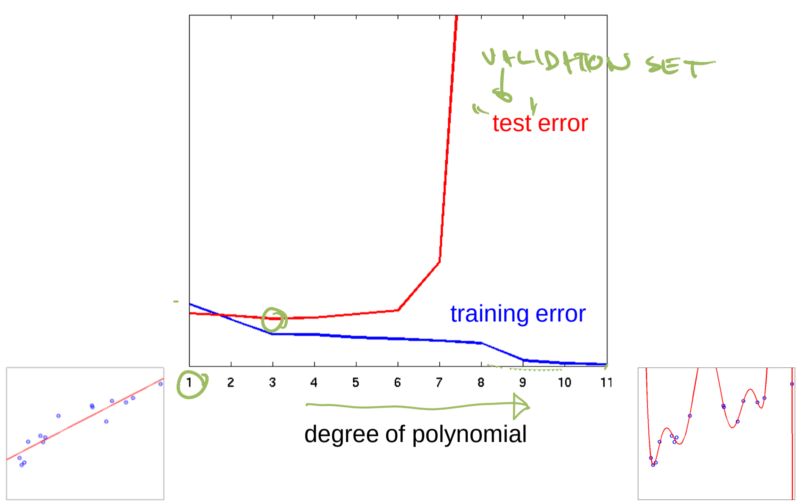
WHAT ABOUT NEW DATA

GENERALIZATION



Training error vs Test error

VALIDATION SET



How to select the model?

FIT MODEL

COMPUTE CURVE

ONLY FOR FINAL MODEL ASSESMENT

Project proposals due next WED

- project title
- project idea
- data set to use
- software to write (if any)
- 1-3 papers to read
- teammates

2-3 people

So far we have covered

Density estimation:

- input: $X_1, X_2, \ldots, X_N$
  output: $p(X)$
  examples: image denoising/segmentation, natural language modeling

Regression:

- input: $(x_1, t_1), (x_2, t_2)\ldots, (x_N, t_N)$
  goal: learn a map $x \mapsto t$
  examples: predicting stock value

Soon to cover

Classification: CATEGORICAL, $\{0,1\}$

- input: $(x_1, y_1), (x_2, y_2)\ldots, (x_N, y_N)$
  goal: learn a map $x \mapsto y$
  examples: object recognition, spam detection

Later in the course

Sequential decisions

- receive observations one at a time
  choose action, receive reward
  examples: driving a vehicle, controling robotic arm