Readings listed in class website

Gaussians
Linear Regression
Bias-Variance Tradeoff

Machine Learning – 10701/15781
Carlos Guestrin
Carnegie Mellon University

September 12th, 2007

What about continuous variables?

- Billionaire says: If I am measuring a continuous variable, what can you do for me?
- **You say: Let me tell you about Gaussians…**

$$P(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

91
92
89
⋮

Some properties of Gaussians

- affine transformation (multiplying by scalar and adding a constant)
  - $X \sim N(\mu,\sigma^2)$
  - $Y = aX + b \rightarrow Y \sim N(a\mu+b,a^2\sigma^2)$
- Sum of Gaussians (independent)
  - $X \sim N(\mu_X,\sigma^2_X)$
  - $Y \sim N(\mu_Y,\sigma^2_Y)$
  - $Z = X+Y \rightarrow Z \sim N(\mu_X+\mu_Y, \sigma^2_X+\sigma^2_Y)$

Learning a Gaussian

- Collect a bunch of data
  - Hopefully, i.i.d. samples
  - e.g., exam scores
- Learn parameters
  - Mean $= E[X] = \frac{1}{n}\sum_i x_i$
  - Variance $= E[(X-\mu)^2]$

$$P(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

MLE for Gaussian

- Prob. of i.i.d. samples $D=\{x_1,\ldots,x_N\}$:

$$P(\mathcal{D} \mid \mu, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{i=1}^{N} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

- Log-likelihood of data:

$$\ln P(\mathcal{D} \mid \mu, \sigma) = \ln\left[\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{i=1}^{N} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}\right]$$

$$= -N \ln \sigma\sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2}$$

Your second learning algorithm: MLE for mean of a Gaussian

- What's MLE for mean?

$$\frac{d}{d\mu} \ln P(\mathcal{D} \mid \mu, \sigma) = \frac{d}{d\mu}\left[-N \ln \sigma\sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$= -\sum_{i=1}^{N} \frac{d}{d\mu}\left[\frac{(x_i-\mu)^2}{2\sigma^2}\right] = \sum_{i=1}^{N} \frac{(x_i-\mu)}{\sigma^2} = 0$$

$$\Rightarrow \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} \mu \Rightarrow \hat{\mu}_{MLE} = \frac{\sum_i x_i}{N}$$

MLE for variance

- Again, set derivative to zero:

$$\frac{d}{d\sigma}\ln P(\mathcal{D}\mid\mu,\sigma) = \frac{d}{d\sigma}\left[-N\ln\sigma\sqrt{2\pi}-\sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$= \frac{d}{d\sigma}\left[-N\ln\sigma\sqrt{2\pi}\right]-\sum_{i=1}^{N}\frac{d}{d\sigma}\left[\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$\frac{-N}{\sigma}+\sum_{i=1}^{N}\frac{2(x_i-\mu)^2}{2\sigma^3}=0$$

$$\Rightarrow \hat{\sigma}^2_{MLE}=\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$$

Learning Gaussian parameters

- MLE:

$$\hat{\mu}_{MLE} = \frac{1}{N}\sum_{i=1}^{N}x_i$$

$$\hat{\sigma}^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(x_i-\hat{\mu}_{MLE})^2$$

true σ^{2*}, then $E_{\mathcal{D}}[\sigma^2_{MLE}] \neq \sigma^{2*}$

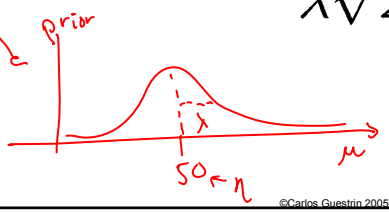
- BTW. MLE for the variance of a Gaussian is **biased**
  - Expected result of estimation is **not** true parameter!
  - Unbiased variance estimator:

$$\hat{\sigma}^2_{unbiased} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\hat{\mu})^2$$

Bayesian learning of Gaussian parameters

- Conjugate priors
  - Mean: Gaussian prior
  - Variance: Wishart Distribution
- Prior for mean:

$$P(\mu \mid \eta, \lambda) = \frac{1}{\lambda\sqrt{2\pi}} e^{\frac{-(\mu-\eta)^2}{2\lambda^2}}$$



MAP for mean of Gaussian

$P(\mu \mid D) \propto P(\mu) \cdot P(D \mid \mu, \sigma)$

prior: $P(\mu \mid \eta, \lambda) = \frac{1}{\lambda\sqrt{2\pi}} e^{\frac{-(\mu-\eta)^2}{2\lambda^2}}$

likelihood: $P(\mathcal{D} \mid \mu, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{i=1}^{N} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$

$$0 = \frac{d}{d\mu}[\ln P(\mathcal{D} \mid \mu) P(\mu)] = \frac{d}{d\mu}[\ln P(\mathcal{D} \mid \mu) + \ln P(\mu)]$$

ln posterior = ln likelihood + ln prior

$$\frac{d}{d\mu} \ln P(D \mid \mu) + \frac{d}{d\mu} \ln P(\mu)$$

$$= \sum_{i=1}^{N} \frac{(x_i-\mu)}{\sigma^2} + \frac{d}{d\mu}\left[\ln \frac{1}{\lambda\sqrt{2\pi}} - \frac{(\mu-\eta)^2}{2\lambda^2}\right]$$

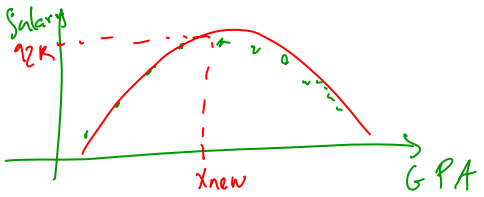
$$= \sum_{i=1}^{N} \frac{(x_i-\mu)}{\sigma^2} - \frac{\mu-\eta}{\lambda^2} = 0$$

$$\sum_{i=1}^{N} \frac{x_i}{\sigma^2} + \frac{\eta}{\lambda^2} = \frac{N\mu}{\sigma^2} + \frac{\mu}{\lambda^2}$$

$$\hat{\mu}_{MAP} = \frac{\sum_{i=1}^{N} \frac{x_i}{\sigma^2} + \frac{\eta}{\lambda^2}}{\frac{N}{\sigma^2} + \frac{1}{\lambda^2}}$$

Prediction of continuous variables

- Billionaire says: Wait, that's not what I meant!
- You says: Chill out, dude.
- He says: I want to predict a continuous variable for continuous inputs: I want to predict salaries from GPA.
- You say: **I can regress that…**



<3.8, 95K>
<4.0, 82K>
<3.7, 100K>
X ; Y

The regression problem

- **Instances:** <x_j, t_j>
- **Learn:** Mapping from x to t(**x**)
- **Hypothesis space:**
  - Given, basis functions
  - Find coeffs **w**={w_1,…,w_k}

$$H = \{h_1, \ldots, h_K\}$$

$$\underbrace{t(\mathbf{x})}_{\text{data}} \approx \hat{f}(\mathbf{x}) = \sum_i w_i h_i(\mathbf{x})$$

  - Why is this called linear regression???
    - model is linear in the parameters

- Precisely, minimize the residual squared error:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_j \left(t(\mathbf{x}_j) - \sum_i w_i h_i(\mathbf{x}_j) \right)^2$$

The regression problem in matrix notation

take derivative and set to 0

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_j \left(t(\mathbf{x}_j) - \sum_i w_i h_i(\mathbf{x}_j) \right)^2$$

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \underbrace{(\mathbf{Hw} - \mathbf{t})^T (\mathbf{Hw} - \mathbf{t})}_{\text{residual error}}$$

$$\mathbf{H} = \begin{bmatrix} h_1 \ldots h_K \\ \vdots \\ \end{bmatrix} \quad \mathbf{w} = \begin{bmatrix} \\ \vdots \\ \end{bmatrix} \quad \mathbf{t} = \begin{bmatrix} t \\ \vdots \\ \end{bmatrix}$$

N samples; K basis func; N sensors

K basis functions | weights | measurements

Regression solution = simple matrix operations

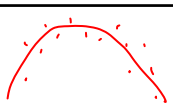
$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \underbrace{(\mathbf{Hw} - \mathbf{t})^T (\mathbf{Hw} - \mathbf{t})}_{\text{residual error}}$$

$$\text{solution: } \mathbf{w}^* = \underbrace{(\mathbf{H}^T\mathbf{H})^{-1}}_{\mathbf{A}^{-1}} \underbrace{\mathbf{H}^T\mathbf{t}}_{\mathbf{b}} = \mathbf{A}^{-1}\mathbf{b}$$

$$\text{where } \mathbf{A} = \mathbf{H}^T\mathbf{H} = \begin{bmatrix} & \\ & \end{bmatrix} \quad \mathbf{b} = \mathbf{H}^T\mathbf{t} = \begin{bmatrix} \\ \\ \\ \end{bmatrix}$$

k×k matrix for k basis functions | k×1 vector

But, why?



$t = X + a$ | $X \sim N(0, \sigma^2)$ | $X+a$ | $t \sim N(a, \sigma^2)$

- Billionaire (again) says: Why sum squared error???
- You say: Gaussians, Dr. Gateson, Gaussians…
- Model: prediction is linear function plus Gaussian noise
  - $t = \sum_i w_i h_i(\mathbf{x}) + \varepsilon$ ← $N(0, \sigma^2)$
  - $P(t \mid x, w, \sigma)$
- Learn **w** using MLE

$$P(t \mid \mathbf{x}, \mathbf{w}, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-[t - \sum_i w_i h_i(\mathbf{x})]^2}{2\sigma^2}}$$

Maximizing log-likelihood

Maximize:

$$\ln P(\mathcal{D} \mid \mathbf{w}, \sigma) = \ln \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{j=1}^{N} e^{\frac{-[t_j - \sum_i w_i h_i(\mathbf{x}_j)]^2}{2\sigma^2}}$$

iid

max

$= \ln \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N - \sum_{j=1}^{N} [t_j - \sum_i w_i h_i(x_j)]^2$

constant — doesn't depend on w

$= \min_w \sum_{j=1}^{N} \left(t_j - \sum_i w_i h_i(x_j)\right)^2$

Least-squares Linear Regression is MLE for Gaussians!!!

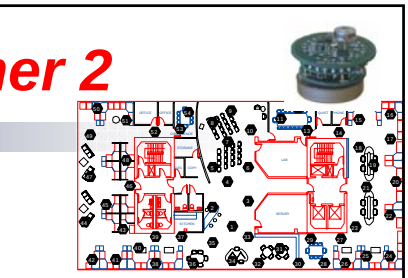
Applications Corner 1

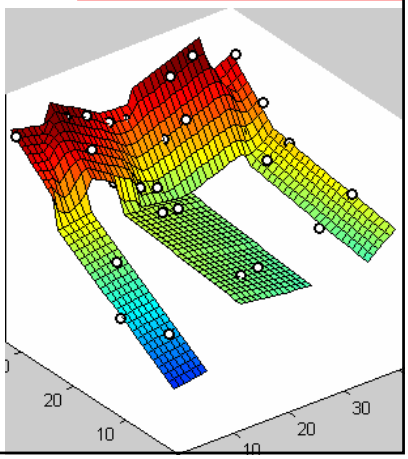
- Predict stock value over time from
  - past values
  - other relevant vars
    - e.g., weather, demands, etc.

17

Applications Corner 2

- Measure temperatures at some locations
- Predict temperatures throughout the environment





[Guestrin et al. '04]

Applications Corner 3

- Predict when a sensor will fail
  - based several variables
    - age, chemical exposure, number of hours used,…

when will sensor fail

19

Announcements 1

- Readings associated with each class
  - See course website for specific sections, extra links, and further details
  - Visit the website frequently
- Recitations
  - Thursdays, 5:00-6:20 in Wean Hall 5409
- Special recitation on Matlab
  - Sept. 18 Tue. 4:30-5:50pm NSH 3002
- Carlos away on Monday Sept. 17th
  - Prof. Eric Xing will teach the lecture

20

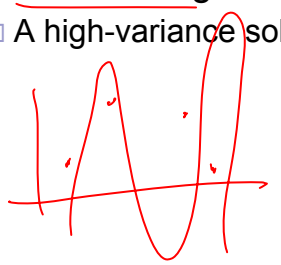
Announcement 2

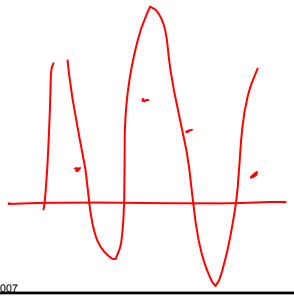
- First homework out later today!
  - Download from course website!
  - Start early!!! :)
  - Due Oct 3rd
- To expedite grading:
  - there are 4 questions
  - please hand in 4 stapled separate parts, one for each question

Bias-Variance tradeoff – Intuition

- Model too "simple" → does not fit the data well
  - A biased solution

- Model too complex → small changes to the data, solution changes a lot
  - A high-variance solution





(Squared) Bias of learner

- Given dataset *D* with *m* samples, learn function h(x)
- If you sample a different datasets, you will learn different h(x)
- **Expected hypothesis:** $E_D[h(x)]$

- **Bias:** difference between what you expect to learn and truth
  - Measures how well you expect to represent true solution
  - Decreases with more complex model

$$bias^2 = \int_x \{E_D[h(x)] - t(x)\}^2 p(x)dx$$

Variance of learner

- Given a dataset *D* with *m* samples, you learn function h(x)
- If you sample a different datasets, you will learn different h(x)
- **Variance:** difference between what you expect to learn and what you learn from a from a particular dataset
  - Measures how sensitive learner is to specific dataset
  - Decreases with simpler model

$$\bar{h}(x) = E_D[h(x)]$$

$$variance = \int E_D[(h(x) - \bar{h}(x))^2]p(x)dx$$

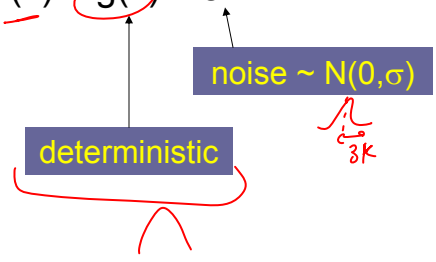
Bias-Variance Tradeoff

- Choice of hypothesis class introduces learning bias
  - More complex class → less bias
  - More complex class → more variance

Bias–Variance decomposition of error

- Consider simple regression problem f:X→T

$t = f(x) = g(x) + \varepsilon$



Collect some data, and learn a function h(x)

What are sources of prediction error?

Sources of error 1 – noise

- What if we have perfect learner, infinite data?
  - If our learning solution h(x) satisfies h(x)=g(x)
  - Still have remaining, *unavoidable error* of σ^2 due to noise ε

$$error(h) = \int_x \int_t (h(x) - t)^2 p(f(x) = t|x) p(x) dt dx$$

$g(x)$ – $g(x)+\varepsilon$

ε^2

variance of ε $= \sigma^2$

27

Sources of error 2 – Finite data

- What if we have imperfect learner, or only m training examples?
- What is our expected squared error per example?
  - Expectation taken over random training sets *D* of size m, drawn from distribution P(X,T)

$$E_D \left[\int_x \int_t \{h(x) - t\}^2 p(f(x) = t|x) p(x) dt dx \right]$$

28

Bias-Variance Decomposition of Error

Bishop Chapter 3

Assume target function: t = f(x) = g(x) + ε

Then expected sq error over fixed size training sets *D* drawn from P(X,T) can be expressed as sum of three components:

$$E_D\left[\int_x\int_t (h(x)-t)^2 p(t|x)p(x)dtdx\right]$$

$$= unavoidableError + bias^2 + variance$$

Where:

$$unavoidableError = \sigma^2$$

$$bias^2 = \int (E_D[h(x)] - g(x))^2 p(x)dx$$

$$\bar{h}(x) = E_D[h(x)]$$

$$variance = \int E_D[(h(x)-\bar{h}(x))^2]p(x)dx$$

What you need to know

- Gaussian estimation
  - MLE
  - Bayesian learning
  - MAP
- Regression
  - Basis function = features
  - Optimizing sum squared error
  - Relationship between regression and Gaussians
- Bias-Variance trade-off
- Play with Applet