***Basics, Gaussians*: Koller&Friedman 1.1, 1.2 – handed out in class**
***Bias-Variance tradeoff*: Bishop chapter 9.1, 9.2**

Gaussians
Linear Regression
Bias-Variance Tradeoff

Machine Learning – 10701/15781

Carlos Guestrin

Carnegie Mellon University

January 23rd, 2006

Announcements

- Recitations stay on Thursdays
  - 5-6:30pm in Wean 5409

- Special Matlab recitation:
  - Jan. 25 Wed. 5:00-7:00pm in NSH 3305

- First homework:
  - Programming part and Analytic part
  - Remember collaboration policy: can discuss questions, but need to write your own solutions and code
  - Out later today
  - Due Mon. Feb 6th **beginning of class**
  - Start early!

Maximum Likelihood Estimation

- **Data:** Observed set *D* of α_H Heads and α_T Tails
- **Hypothesis:** Binomial distribution
- Learning θ is an optimization problem
  - What's the objective function?

- MLE: Choose θ that maximizes the probability of observed data:

$$\begin{aligned}
\widehat{\theta} &= \arg\max_{\theta} \; P(\mathcal{D} \mid \theta) \\
&= \arg\max_{\theta} \; \ln P(\mathcal{D} \mid \theta)
\end{aligned}$$

Bayesian Learning for Thumbtack

$$P(\theta \mid \mathcal{D}) \;\propto\; P(\mathcal{D} \mid \theta)P(\theta)$$

- Likelihood function is simply Binomial:

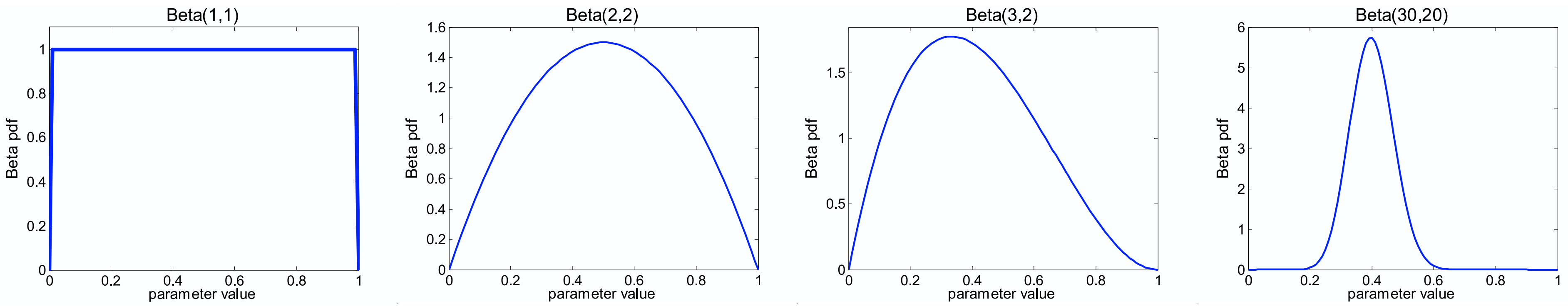
$$P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

- What about prior?
  - Represent expert knowledge
  - Simple posterior form
- Conjugate priors:
  - Closed-form representation of posterior
  - **For Binomial, conjugate prior is Beta distribution**

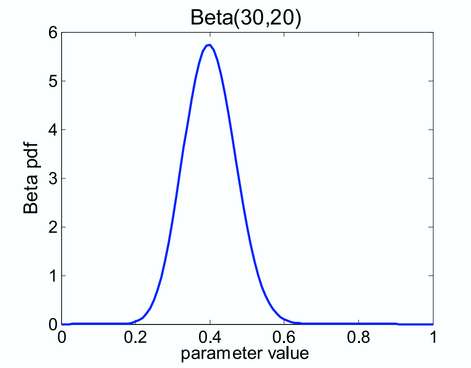
Posterior distribution

- Prior: $Beta(\beta_H, \beta_T)$
- Data: α_H heads and α_T tails
- Posterior distribution:

$$P(\theta \mid \mathcal{D}) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$



MAP: Maximum a posteriori approximation



$$P(\theta \mid \mathcal{D}) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

$$E[f(\theta)] = \int_0^1 f(\theta) P(\theta \mid \mathcal{D}) d\theta$$

- As more data is observed, Beta is more certain

- MAP: use most likely parameter:

$$\widehat{\theta} = \arg\max_{\theta} P(\theta \mid \mathcal{D}) \qquad E[f(\theta)] \approx f(\widehat{\theta})$$

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…**

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

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
  - $X \sim N(\mu_X,\sigma^2_X)$
  - $Y \sim N(\mu_Y,\sigma^2_Y)$
  - $Z = X+Y \rightarrow Z \sim N(\mu_X+\mu_Y, \sigma^2_X+\sigma^2_Y)$

Learning a Gaussian

- Collect a bunch of data
  - Hopefully, i.i.d. samples
  - e.g., exam scores
- Learn parameters
  - Mean
  - Variance

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

MLE for Gaussian

- Prob. of i.i.d. samples $x_1,\ldots,x_N$:

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

- Log-likelihood of data:

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

Your second learning algorithm: MLE for mean of a Gaussian

- What's MLE for mean?

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

MLE for variance

- Again, set derivative to zero:

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

Learning Gaussian parameters

- MLE:

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

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

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

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

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

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

$$\boxed{\frac{d}{d\mu}\left[\ln P(\mathcal{D} \mid \mu) P(\mu)\right] \;=\; \frac{d}{d\mu}\left[\ln P(\mathcal{D} \mid \mu) + \ln P(\mu)\right]}$$

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…**

The regression problem

- **Instances: <$\mathbf{x}_j$, t_j>**
- **Learn:** Mapping from x to t(**x**)
- **Hypothesis space:**
  - Given, basis functions $H = \{h_1, \ldots, h_K\}$
  - Find coeffs **w**={w_1,…,w_k} $\underbrace{t(\mathbf{x})}_{\text{data}} \approx \widehat{f}(\mathbf{x}) = \sum_i w_i h_i(\mathbf{x})$
  - Why is this called linear regression???
    - model is linear in the parameters

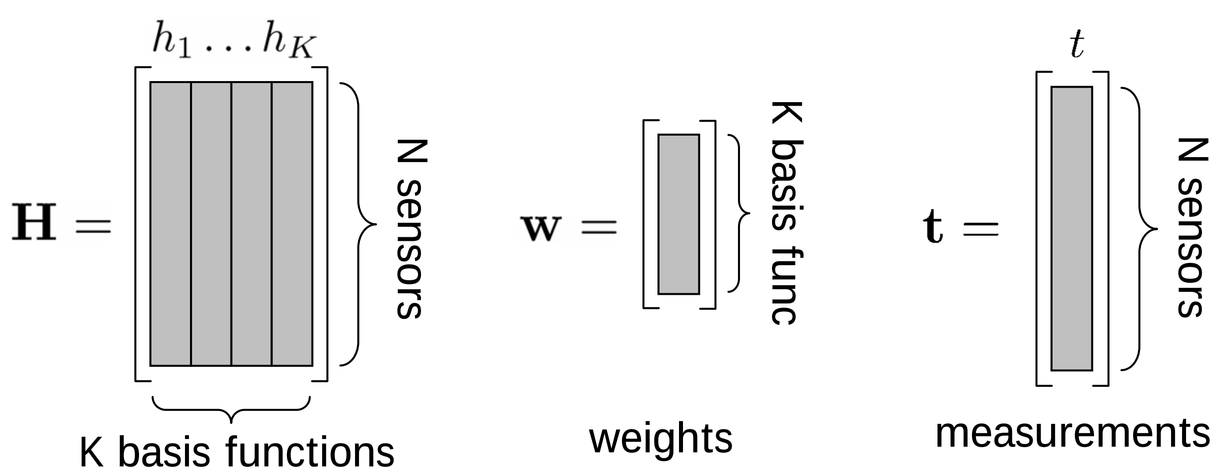
- Precisely, minimize the residual error:

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

The regression problem in matrix notation

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

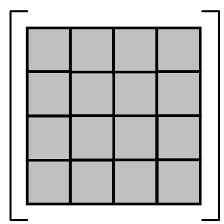
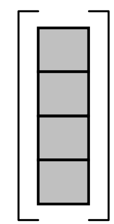
$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \underbrace{(\mathbf{H}\mathbf{w} - \mathbf{t})^T (\mathbf{H}\mathbf{w} - \mathbf{t})}_{\text{residual error}}$$



Regression solution = simple matrix operations

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

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

where $\mathbf{A} = \mathbf{H}^{\mathrm{T}}\mathbf{H} =$  $\mathbf{b} = \mathbf{H}^{\mathrm{T}}\mathbf{t} =$ 

k×k matrix
for k basis functions

k×1 vector

But, why?

- Billionaire (again) says: Why sum squared error???
- You say: Gaussians, Dr. Gateson, Gaussians…

- Model: prediction is linear function plus Gaussian noise
  - $t = \sum_i w_i h_i(\mathbf{x}) + \varepsilon$

- Learn **w** using MLE

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

Maximizing log-likelihood

Maximize:

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

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

Bias-Variance tradeoff – Intuition

- Model too "simple" $\rightarrow$ does not fit the data well
  - A biased solution

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

(Squared) Bias of learner

- Suppose you are given a dataset *D* with *m* samples from some distribution
- You learn function h(x) from data *D*
- 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

- Suppose you are given a dataset *D* with *m* samples from some distribution
- You learn function h(x) from data *D*
- 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

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

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

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

  noise ~ $N(0,\sigma)$

  deterministic

  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?
  - Our learning solution h(x) satisfies h(x)=g(x)
  - Still have remaining, *unavoidable error* of σ^2 due to noise ε

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

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)

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

Bias-Variance Decomposition of Error

Bishop chapter 9.1, 9.2

Assume target function: $t = f(x) = g(x) + \varepsilon$

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