CS540 Machine learning
Lecture 2

Announcements

- Matlab tutorial by Ian Mitchell
  today Thursday Sept 11, 5pm-7pm, Dmp 101
- Don’t send me typos by email! Instead follow the procedure described here

http://www.cs.ubc.ca/~murphyk/MLbook/typos.html

Use pdf page numbers (= book + 24)

- Changed order of topics (spiral method)

Outline

- Your to-do list for today
- Quiz
- Data
- Probabilistic Models
- Maximum likelihood estimation

Your to-do list

- Buy textbook
- Read ch 1-2
- Join google groups (so far 21 members)
- Get access to matlab 2008a
- Attend matlab tutorial tonight (beginners)
- Work through Matt Dunham's matlab tutorial (beginners and experts)
- Do homework 1 (due Tuesday)

Quiz

- About half of you did fine, other half need to revise their maths
- If you can't / don't want to do math, don't take this class. CS340 may be a better bet.
- One person wrote "I know the answer intuitively, the proof doesn't matter… I want to learn ML deeply" – this is a contradiction.

Q1.1

- A rotation in 3d by angle α about the z axis is given by the following matrix:

$$\mathbf{R}(\alpha) = \begin{pmatrix} \cos(\alpha) & -\sin(\alpha) & 0 \\ \sin(\alpha) & \cos(\alpha) & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

- Prove that R is an orthogonal matrix, i.e., R'R=I, for any α.
- Most people got this fine. Need to remember that

$$\cos^2 + \sin^2 = 1$$

Q1.2

- A rotation in 3d by angle α about the z axis is given by the following matrix:

$$\mathbf{R}(\alpha) = \begin{pmatrix} \cos(\alpha) & -\sin(\alpha) & 0 \\ \sin(\alpha) & \cos(\alpha) & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

- What is the only eigenvector v of R with an eigenvalue of 1.0?
- I forgot to add constraint that v is of unit norm, $||v||^2=1$
- With constraint, v = (0,0,1) or v=(0,0,-1) are only valid solutions (axis of rotation!)
- Without constraint, v=(0,0,z) is a set of solutions.

Q1.2

- Solving

$$\begin{pmatrix} c & s & 0 \\ -s & c & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}, \quad x^2 + y^2 + z^2 = 1$$

- Gives x=0,y=0,z=1 or z=-1
- Using symbolic math toolbox

```
syms c s x y z
S=solve('c*x-s*y=x','s*x+c*y=y','x^2+y^2+z^2=1')
>> S.x = [0 0], S.y = [0 0], S.z = [1 -1]
```

Q2.1

- Derivatives – most people got this

$$\frac{\partial}{\partial x_j} \sum_i a_i x_i = a_j, \quad \frac{\partial(\mathbf{a}^T \mathbf{x})}{\partial \mathbf{x}} = \mathbf{a}$$

Q3.1

- You will soon know this off by heart

$$\mathcal{N}(x|\mu,\sigma^2) \stackrel{\text{def}}{=} \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{1}{2\sigma^2}(x-\mu)^2]$$

Q3.2

$$\mathcal{N}(x|\mu,\sigma^2) \stackrel{\text{def}}{=} \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{1}{2\sigma^2}(x-\mu)^2]$$

- Densities can have p(x)>1 so long as

$$\int p(x)dx = 1$$

- Eg narrow gaussian density at its mode

$$\mathcal{N}(\mu|\mu,\sigma^2) = (\sigma\sqrt{2\pi})^{-1}e^0$$

If $\sigma < 1/\sqrt{2\pi}$, we have $p(x) > 1$

```
s=1/sqrt(2*pi); normpdf(0,0,0.9*s)
```

Q3.3

- X in {0,1} so

$$EX = \sum_{x \in \{0,1\}} xp(x) = 0 \times p(x=0) + 1 \times p(x=1) = \theta$$

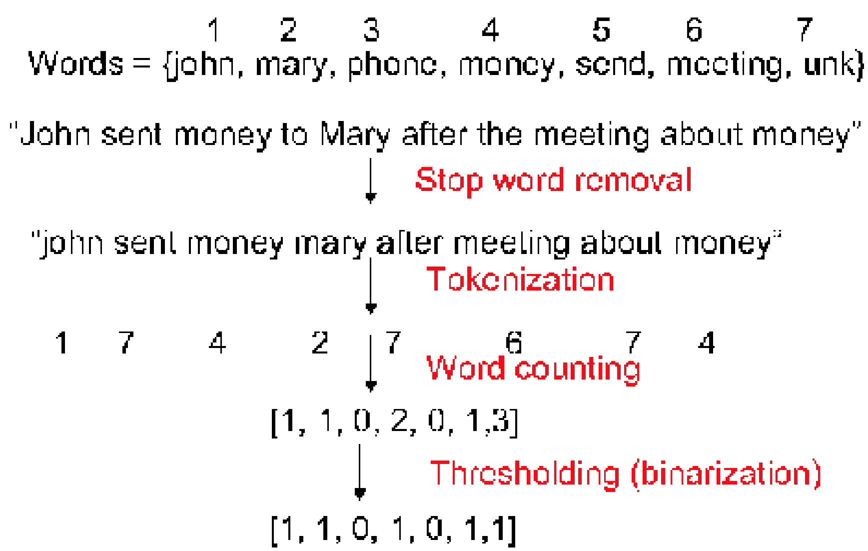
- Similarly for variance.
- Many people used an integral instead of a sum…

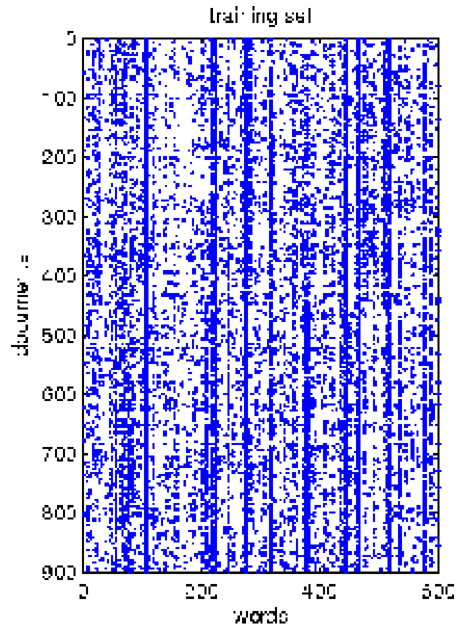
Outline

- Your to-do list for today
- Quiz
- Data
- Probabilistic Models
- Maximum likelihood estimation

Feature vectors

- Often need to convert data into fixed-length feature vectors, so X is n x d design matrix.
- Eg bag-of-words representation of text documents





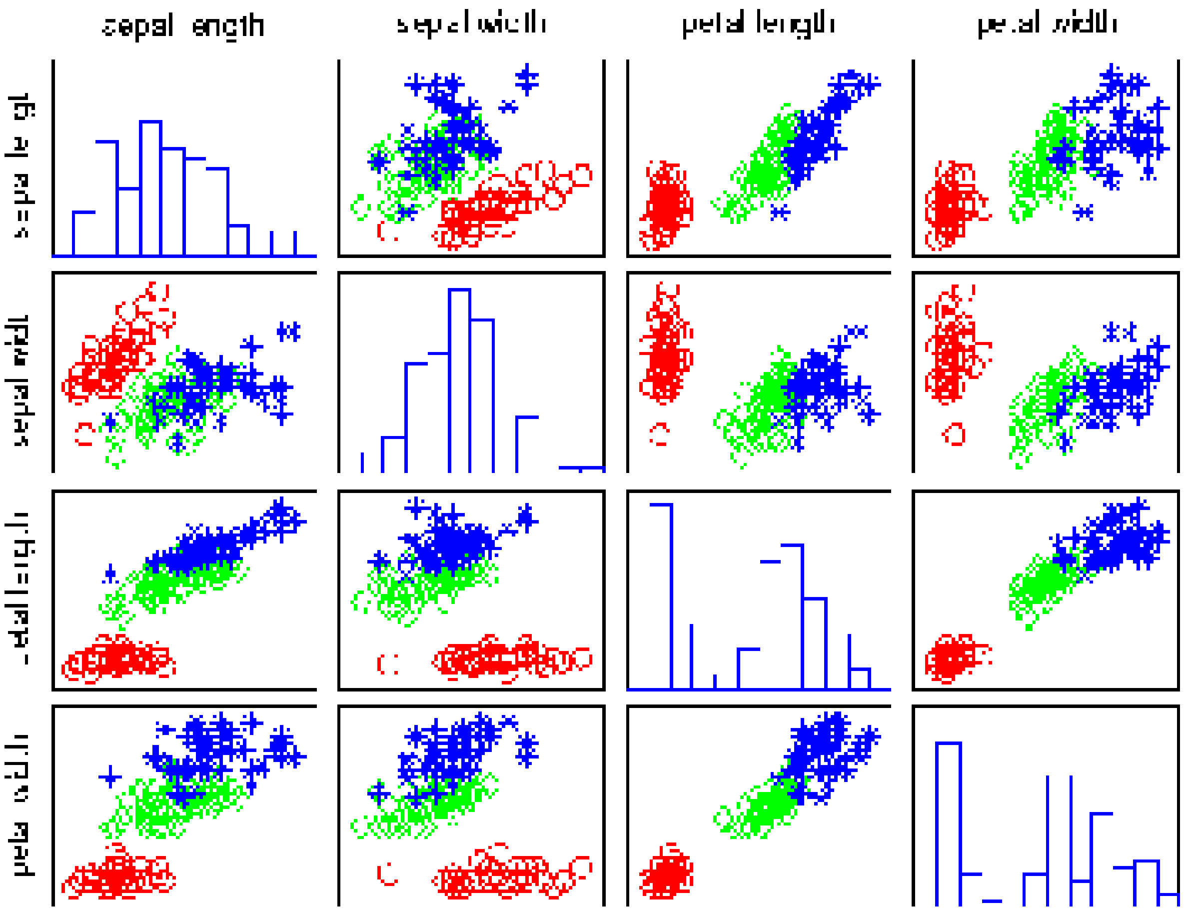
Design matrix

- Sometimes features given to us (Fisher's iris data)



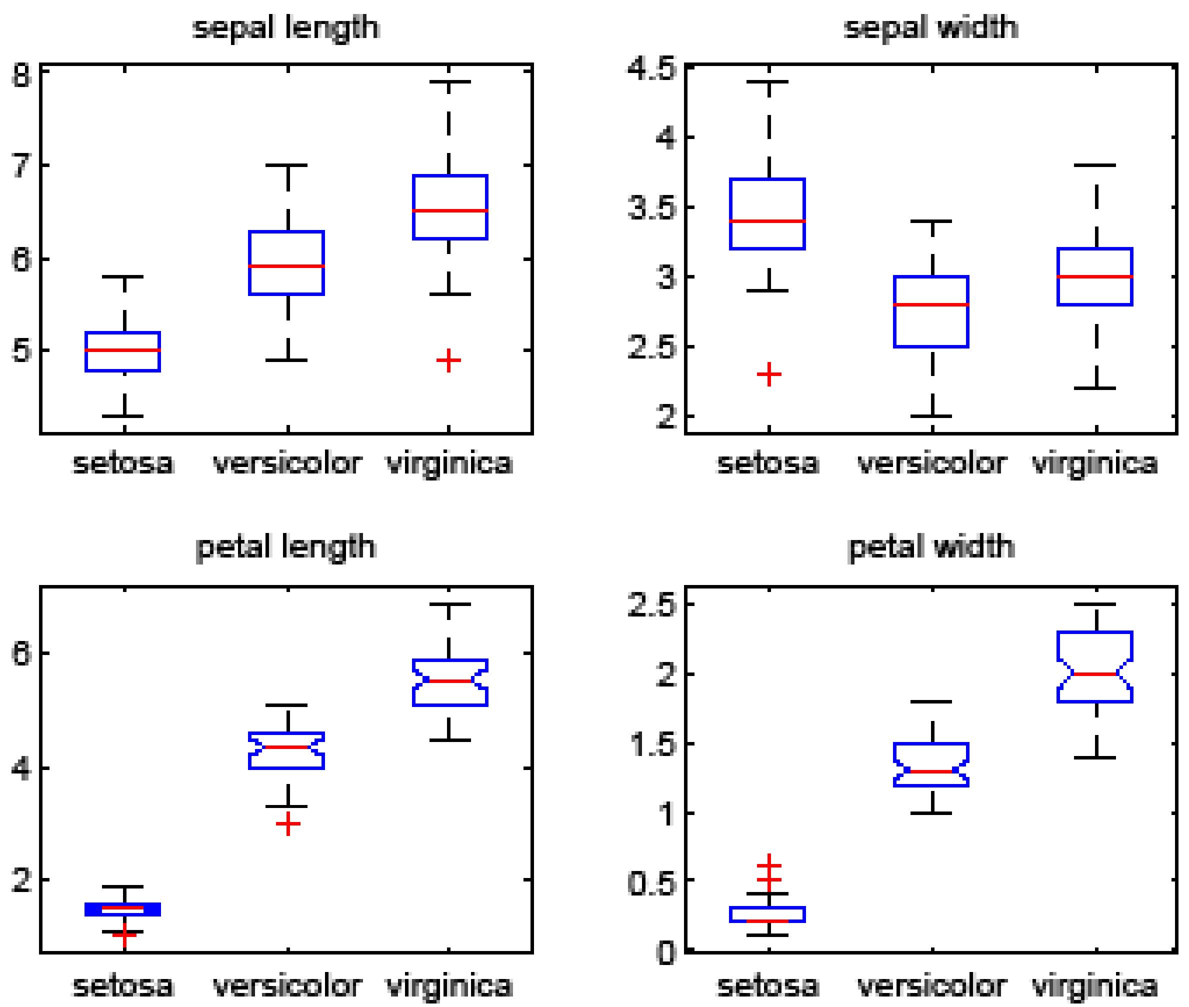
Case	Sepal length	Sepal width	Petal length	Petal width	Class
1	5.1000	3.5000	1.4000	0.2000	setosa
2	4.9000	3.0000	1.4000	0.2000	setosa
			⋮		
51	7.0000	3.2000	4.7000	1.4000	versicolor
52	6.4000	3.2000	4.5000	1.5000	versicolor
			⋮		
101	6.3000	3.3000	6.0000	2.5000	virginica
			⋮		
150	5.9000	3.0000	5.1000	1.8000	virginica

Pairwise scatter plot



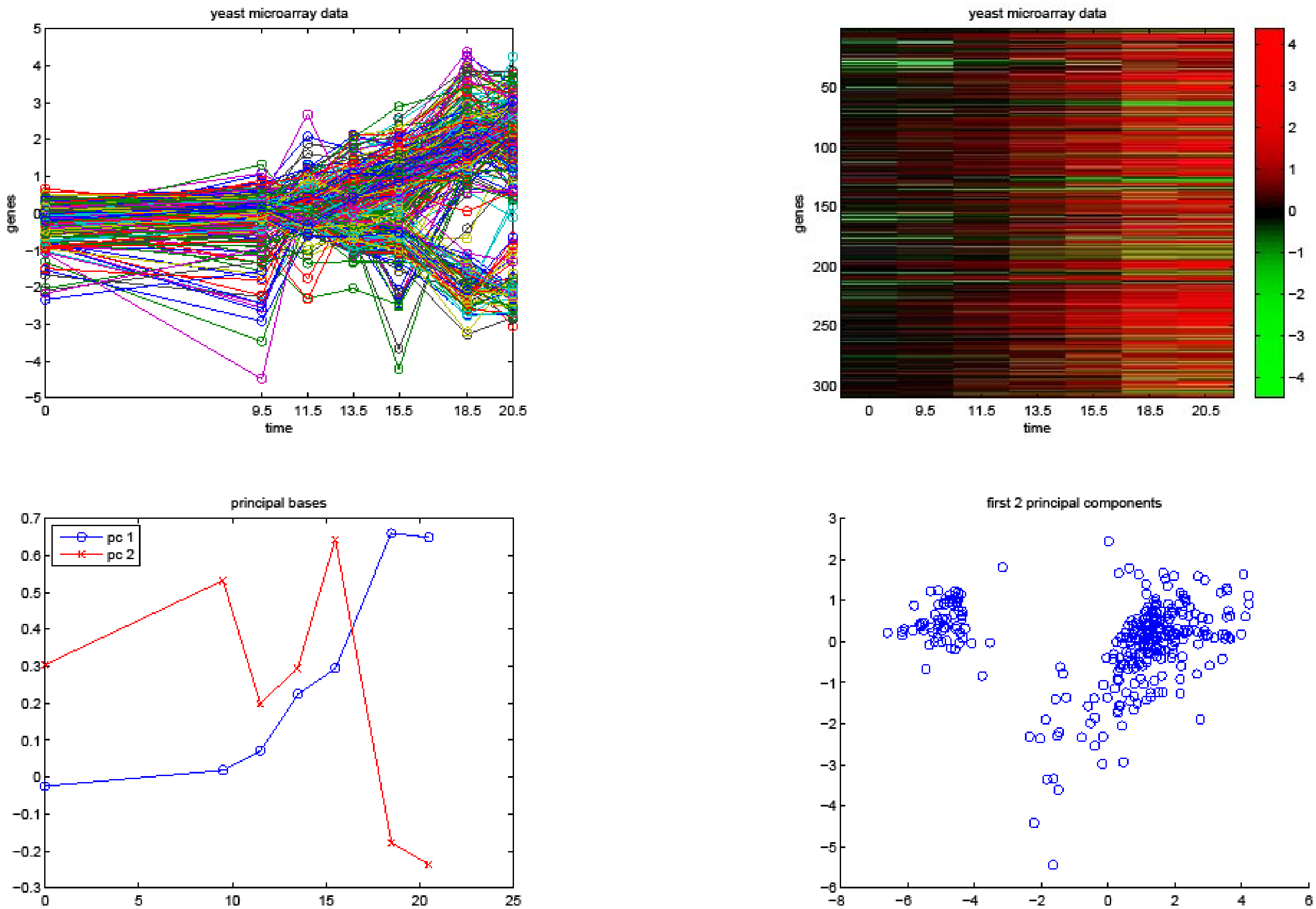
pscatter(X,’y’,y) in matlab (MLAPA function)

Boxplots



sepal length
8
7
6
5
setosa
versicolor
virginica
sepal width
4.5
4
3.5
3
2.5
2
setosa
versicolor
virginica
petal length
6
4
2
setosa
versicolor
virginica
petal width
2.5
2
1.5
1
0.5
0
setosa
versicolor
virginica

Visualizing data using PCA



Principal components analysis

Outline

- Your to-do list for today
- Quiz
- Data
- Probabilistic Models
- Maximum likelihood estimation

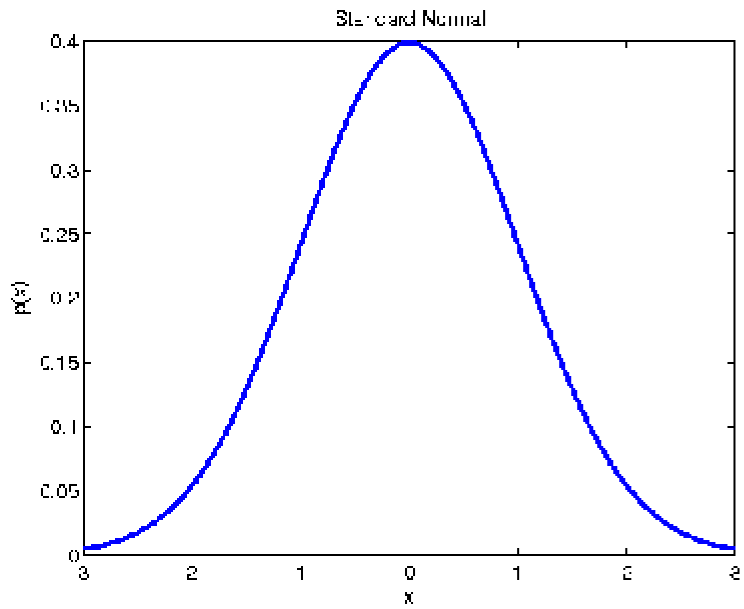
Models

- (Univariate) Gaussian
- Bernoulli/ Binomial
- Multinomial
- Linear regression
- Logistic regression

The Gaussian (normal)

- Most widely used distribution (central limit theorem, maxent, mathematical tractability)

$$\mathcal{N}(x|\mu,\sigma^2) \stackrel{\text{def}}{=} \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{1}{2\sigma^2}(x-\mu)^2]$$



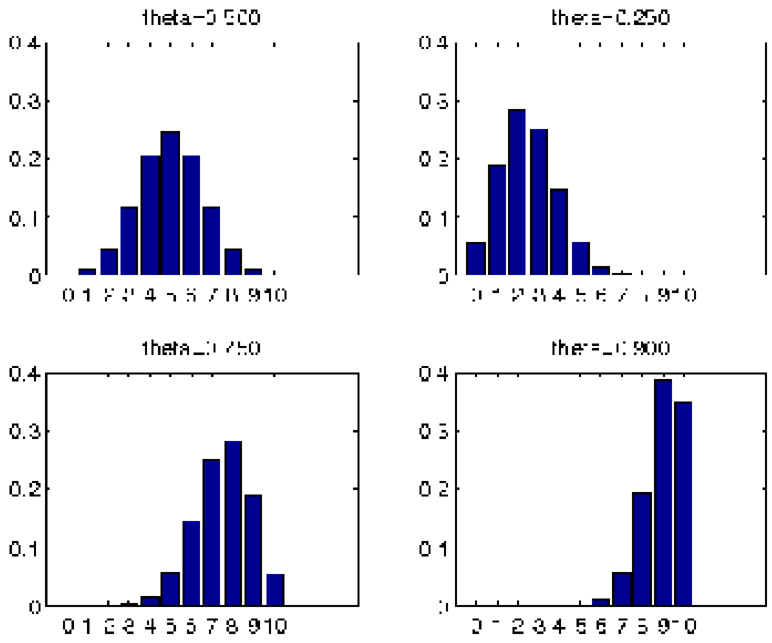
Bernoulli/Binomial

- Popular model for binary data eg. coin tossing

$$\mathrm{Ber}(x|\theta) \stackrel{\mathrm{def}}{=} \theta^x (1-\theta)^x = \theta^{I(x=1)}(1-\theta)^{I(x=0)}$$

- Binomial

$$\mathrm{Bin}(x|\pi, m) \stackrel{\mathrm{def}}{=} \binom{m}{x} \pi^x (1-\pi)^{m-x}$$



Multinomial

- Coins to dice

$$\mathrm{Mu}(\mathbf{x}|m, \boldsymbol{\pi}) = \binom{m}{x_1 \ldots x_K} \prod_{j=1}^{K} \pi_j^{x_j}$$

$$\binom{m}{x_1 \ldots x_K} = \frac{m!}{x_1! x_2! \cdots x_K!}$$

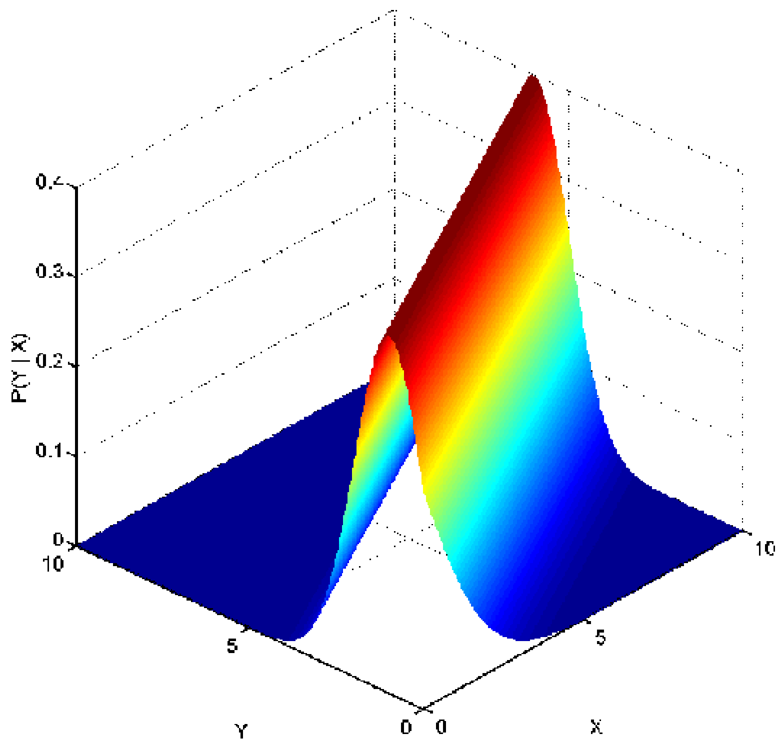
- If m=1, we get 1-of-K encoding of categorical variable

$$p(\mathbf{x}|\boldsymbol{\pi}) = \mathrm{Mu}(\mathbf{x}|\boldsymbol{\pi}, 1) = \prod_{j=1}^{K} \pi_j^{I(x_j=1)}$$

Linear regression

- Gaussian is unconditional density p(y)
- Linear regression is conditional density p(y|x)

$$\begin{aligned} p(y|\mathbf{x}, \boldsymbol{\theta}) &= \mathcal{N}(y|\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}), \sigma^2) \\ y &= \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}) + \epsilon, \ \ \epsilon \sim \mathcal{N}(0, \sigma^2) \end{aligned}$$

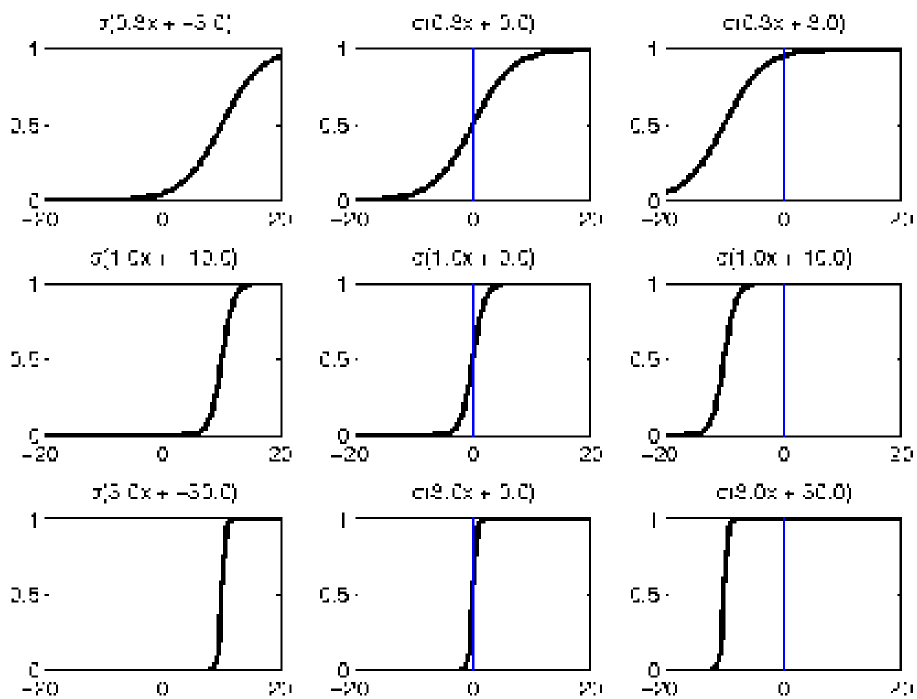


Logistic regression

- Model for binary *classification*

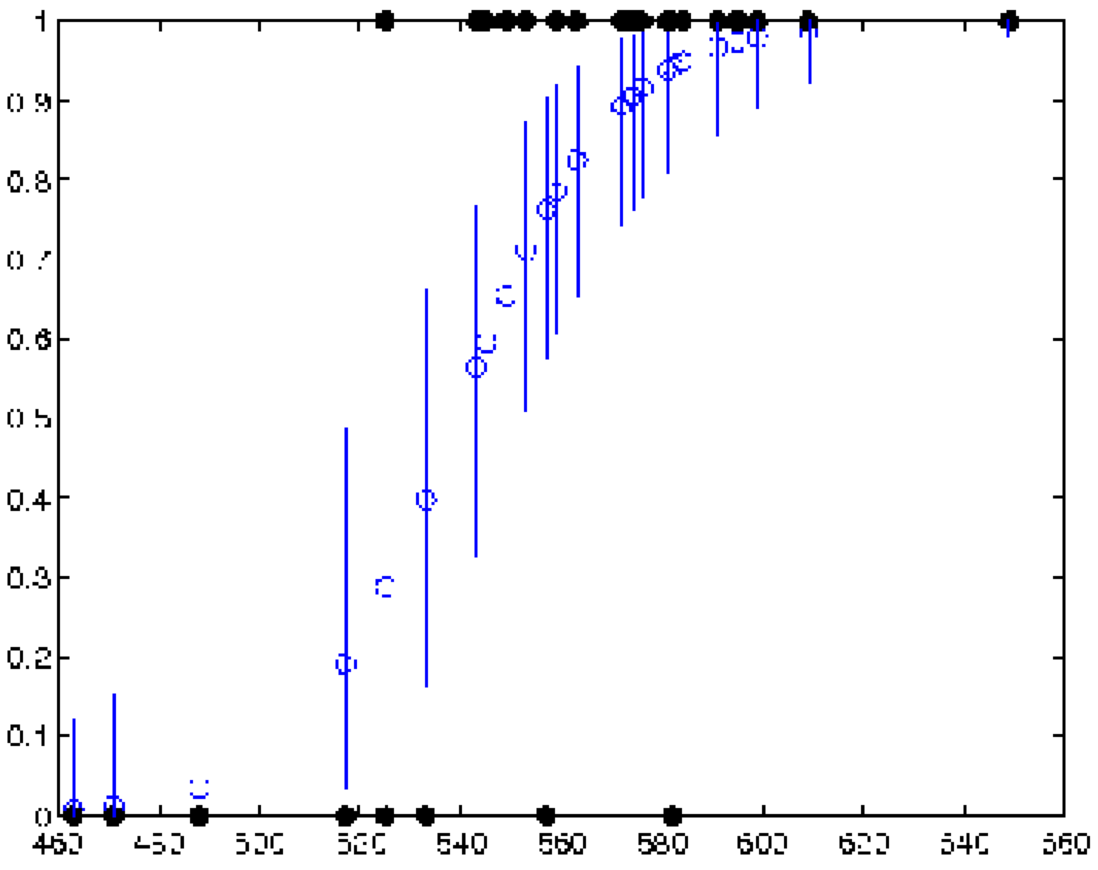
$$p(y=1|\mathbf{x},\mathbf{w}) \quad = \quad \sigma(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x})) = \sigma(\sum_{j=1}^{d} w_j\phi_j)$$

$$\sigma(\eta) \quad \stackrel{\text{def}}{=} \quad \frac{1}{1+\exp(-\eta)} = \frac{e^{\eta}}{e^{\eta}+1}$$



Sigmoid
or
Logistic
function

Logistic regression



SAT scores

Outline

- Your to-do list for today
- Quiz
- Data
- Probabilistic Models
- Maximum likelihood estimation

Maximum likelihood estimation (MLE)

$$\hat{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \arg\max_{\boldsymbol{\theta}} p(\mathcal{D}|\boldsymbol{\theta})$$

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{x}_i|\boldsymbol{\theta})$$

exchangeability

$$\ell(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \log p(\mathcal{D}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(\mathbf{x}_i|\boldsymbol{\theta})$$

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$$

monotonocity of log

MLE for Gaussian mean

$$
\begin{aligned}
\ell(\mu, \sigma^2) &= \sum_{i=1}^{n} \log \mathcal{N}(x_i|\mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 - \frac{n}{2} \ln \sigma^2 - \frac{n}{2} \ln(2\pi) \\
\frac{\partial \ell}{\partial \mu} &= -\frac{2}{2\sigma^2} \sum_i (x_i - \mu) = 0 \\
\hat{\mu} &= \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{x}
\end{aligned}
$$

MLE = empirical mean

MLE for Gaussian variance

$$
\begin{aligned}
\ell(\mu, \sigma^2) &= \sum_{i=1}^{n} \log \mathcal{N}(x_i|\mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 - \frac{n}{2} \ln \sigma^2 - \frac{n}{2} \ln(2\pi) \\
\frac{\partial \ell}{\partial \sigma^2} &= 0 \\
\hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 \\
&= \left(\frac{1}{n} \sum_i x_i^2 \right) - (\overline{x})^2
\end{aligned}
$$

See book for algebra

MLE vs unbiased estimator

- MLE

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

var(x,1)

- Unbiased estimator

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

var(x)

- See sec 11.4

MLE for Bernoulli/Binomial

$$\mathrm{Ber}(x|\theta) \stackrel{\mathrm{def}}{=} \theta^x (1-\theta)^x = \theta^{I(x=1)} (1-\theta)^{I(x=0)}$$

- MLE

$$\begin{aligned} p(D|\theta) &= \prod_{i=1}^{n} p(x_i|\theta) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} = \theta^{N_1} (1-\theta)^{N_0} \\ \ell(\theta) &= \log p(D|\theta) = N_1 \log \theta + N_0 \log(1-\theta) \\ \hat{\theta} &= \frac{N_1}{n} \end{aligned}$$

MLE for multinomial

- Log likelihood

$$\ell = \sum_k N_k \log \pi_k$$

- Need to enforce sum-to-one constraint with Lagrange multiplier

$$\tilde{\ell} = \sum_k N_k \log \pi_k + \lambda \left(1 - \sum_k \pi_k \right)$$

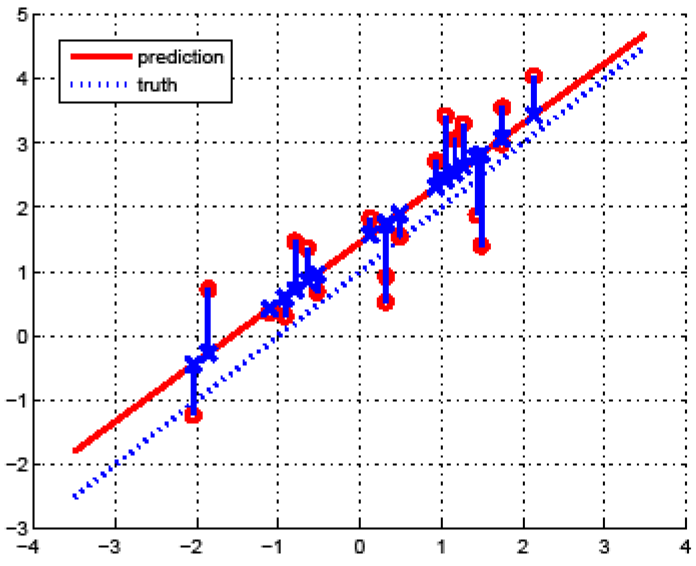
- Simple algebra yields

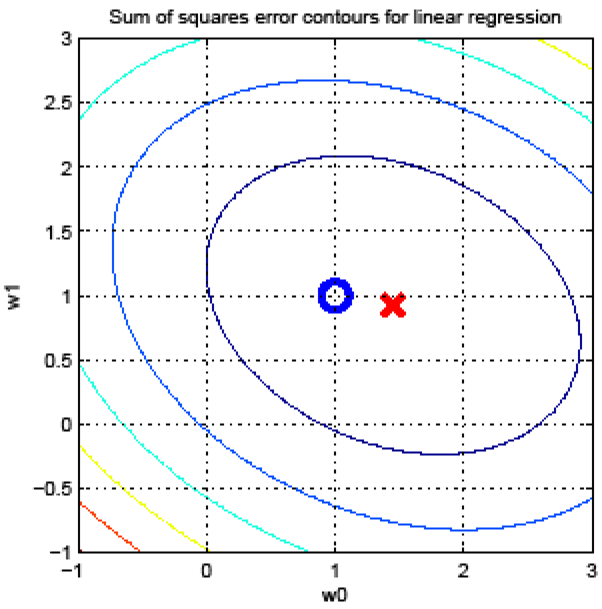
$$\hat{\pi}_k = \frac{N_k}{N}$$

MLE for linear regression (least squares)

$$
\begin{aligned}
p(\mathcal{D}|\mathbf{w},\sigma^2) &= \prod_{i=1}^{n} \mathcal{N}(y_i|\mathbf{w}^T\mathbf{x}_i,\sigma^2) \\
&= \prod_{i=1}^{n} (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\mathbf{w})^T(\mathbf{y}-\mathbf{X}\mathbf{w})\right) \\
J(\mathbf{w},\sigma^2) &= -\log p(\mathbf{y}|X,\mathbf{w},\sigma^2) \\
&= \frac{n}{2}\log(\sigma^2) + \frac{1}{2\sigma^2} RSS(\mathbf{w}) \\
RSS(\mathbf{w}) &= ||\mathbf{X}\mathbf{w}-\mathbf{y}||_2^2 = (\mathbf{y}-\mathbf{X}\mathbf{w})^T(\mathbf{y}-\mathbf{X}\mathbf{w})
\end{aligned}
$$

Negative log likelihood





Normal equations

$$\begin{aligned}\nabla_{\mathbf{w}} RSS(\mathbf{w}) &= \mathbf{0} \\ \hat{\mathbf{w}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = (\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T)^{-1}(\sum_{i=1}^{n} y_i \mathbf{x}_i)\end{aligned}$$

See book for derivation

MLE = OLS estimate

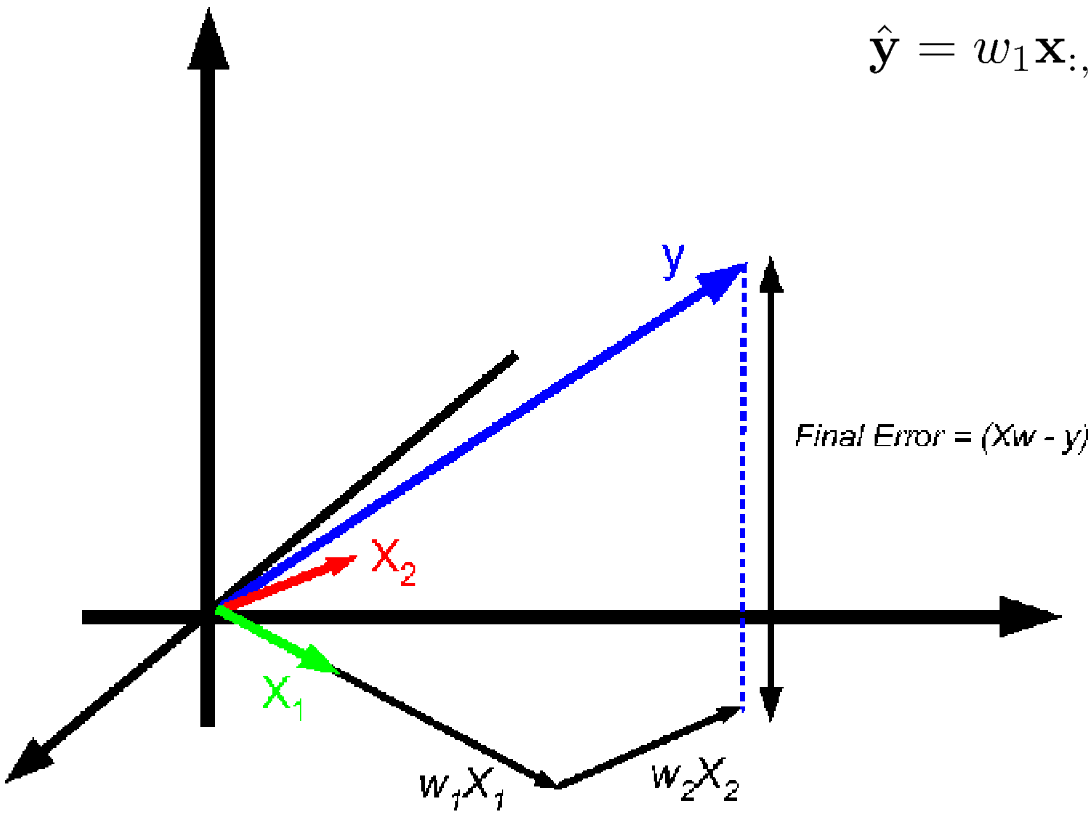
Uncertainty in estimate – see later

Geometry of least squares

$$\mathbf{X} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \\ 0 & 0 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 3 \\ 2 \\ 3 \end{pmatrix}$$

Minimize RSS by orthogonal projection of y into column space of X

$$\hat{\mathbf{y}} = w_1 \mathbf{x}_{:,1} + \cdots + w_d \mathbf{x}_{:,d}$$



Orthogonal projection

- Prediction on the training set

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{w}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \stackrel{\text{def}}{=} \mathbf{H}\mathbf{y}$$

- Residual error is orthogonal to X

$$\mathbf{X}^T(\mathbf{y} - \mathbf{H}\mathbf{y}) = \mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}) = \mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{0}$$

Logistic regression

- Likelihood has unique global maximum (you will prove this in a later homework)
- But MLE has no closed form solution
- We will discuss algorithms later