ECE521 W17 Tutorial 6

Min Bai and Yuhuai (Tony) Wu

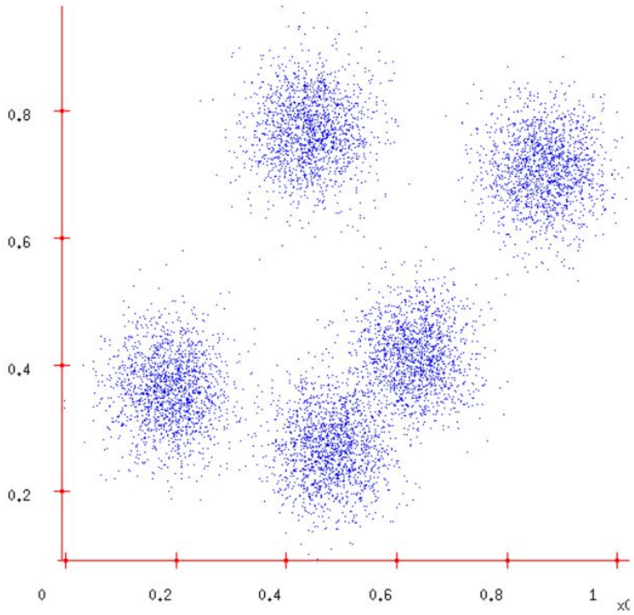
UNIVERSITY OF TORONTO

Agenda

- kNN and PCA
- Bayesian Inference

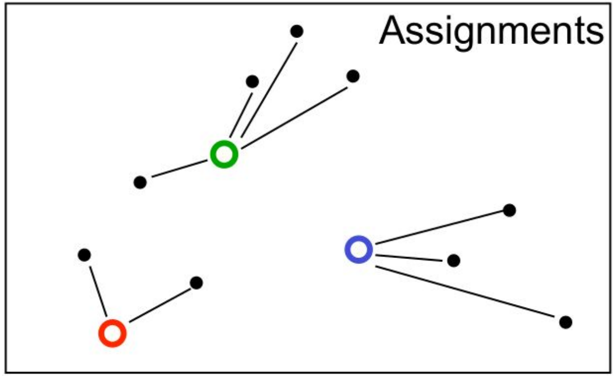
k-Means

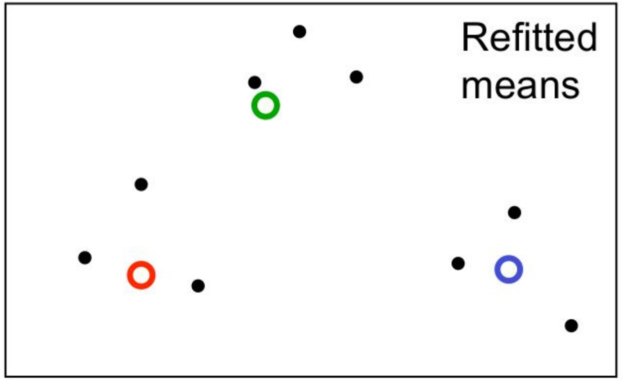
- Technique for clustering
- Unsupervised pattern and grouping discovery
- Class prediction
- Outlier detection



k-Means

- Assume the data lives in a Euclidean space.
- Assume we want k classes/patterns
- Initialization: randomly located cluster centers
- The algorithm alternates between two steps:
  - Assignment step: Assign each datapoint to the closest cluster.
  - Refitting step: Move each cluster center to the center of gravity of the data assigned to it.





k-Means

- Define an iterative procedure to minimize:

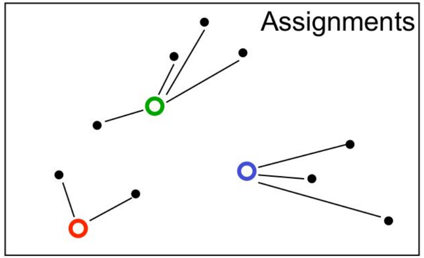
$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||\mathbf{x}_n - \boldsymbol{\mu}_k||^2.$$

- Given $\boldsymbol{\mu}_k$, minimize J with respect to r_{nk} (akin to the **E-step** in EM):

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j ||\mathbf{x}_n - \boldsymbol{\mu}_j||^2 \\ 0 & \text{otherwise} \end{cases}$$

Hard assignments of points to clusters

which simply says assign the n^{th} data point $\mathbf{x}_n$ to its closest cluster centre

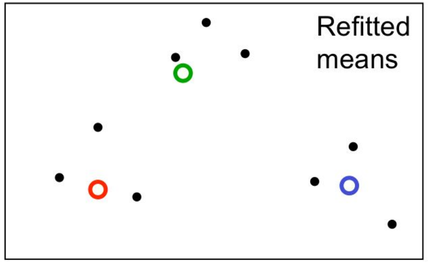


- Given r_{nk}, minimize J with respect to $\boldsymbol{\mu}_k$ (akin to the **M-step**):

$$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}}.$$

Number of points assigned to cluster k.

Set $\boldsymbol{\mu}_k$ equal to the mean of all the data points assigned to cluster k

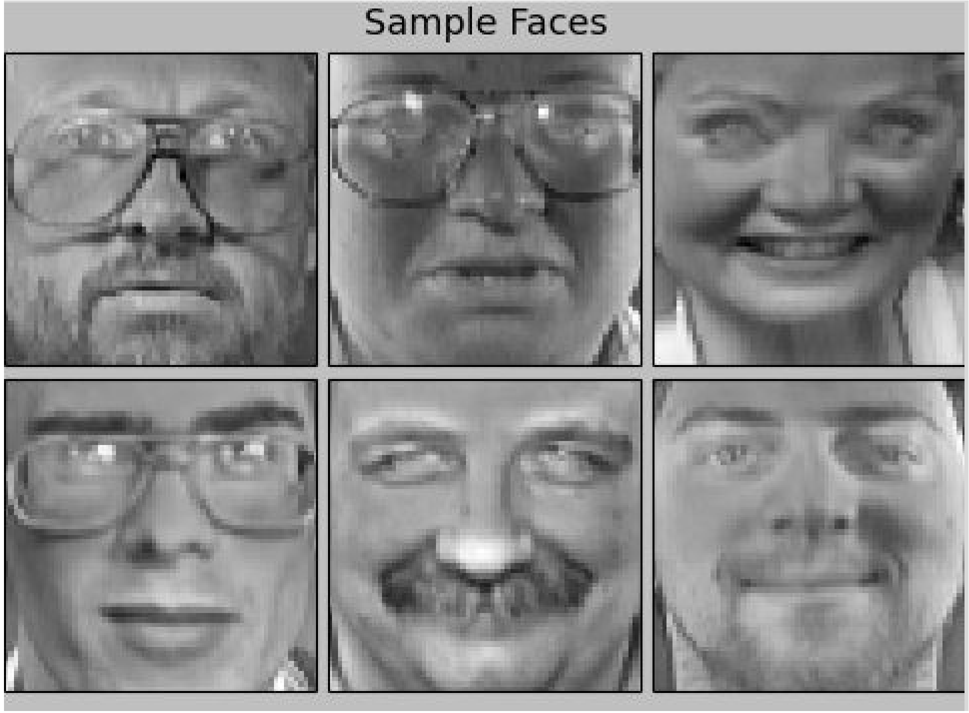


PCA

- Dimensionality Reduction
- Visualization
- Compression

Olivetti Faces Dataset

- Gray scale images
- 64x64
- 10 subjects
- 40 images per subject
- 400 images each
- Problem: 4096-dim feature vector for only 400 datapoints
- Would like: 200 dimensional feature space (each image described by 200 numbers)



PCA

- Algorithm: to find M components underlying D-dimensional data
  1. Select the top M eigenvectors of C (data covariance matrix):

$$C = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}^{(n)} - \bar{\mathbf{x}})(\mathbf{x}^{(n)} - \bar{\mathbf{x}})^T = U\Sigma U^T \approx U\Sigma_{1:M}U_{1:M}^T$$

     where U: orthogonal, columns = unit-length eigenvectors

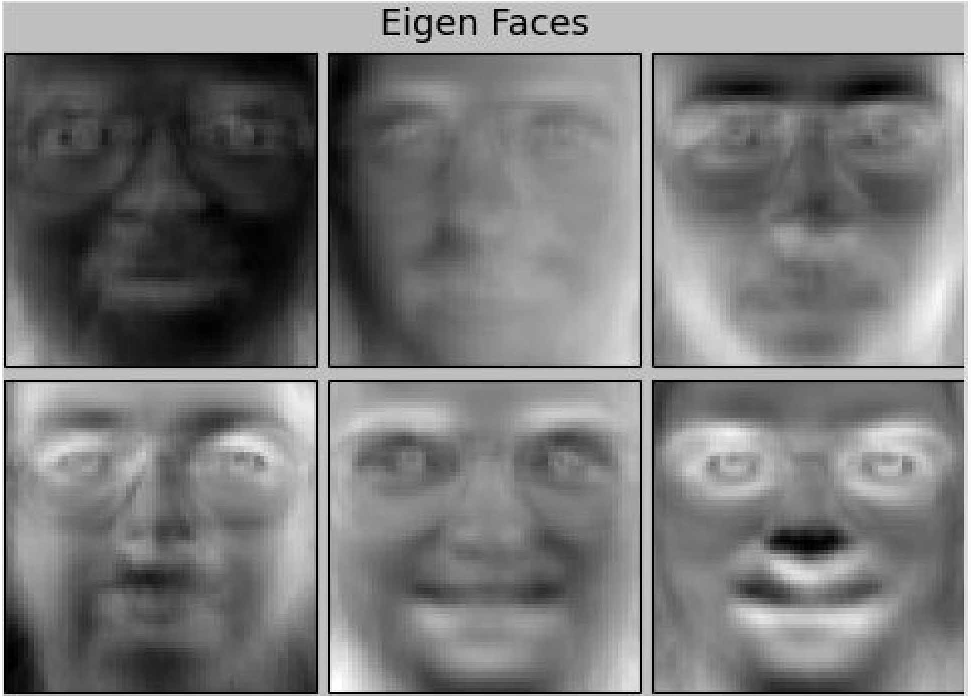
$$U^T U = UU^T = 1$$

     and Σ: matrix with eigenvalues in diagonal = variance in direction of eigenvector
  2. Project each input vector x into this subspace, e.g.,

$$z_j = \mathbf{u}_j^T\mathbf{x}; \qquad \mathbf{z} = U_{1:M}^T\mathbf{x}$$

Olivetti Faces Dataset - PCA

- First six principal components (eigen faces) u_0, … u_5
- u_j is column j of matrix U



PCA Reconstruction

- Z^n is the list of coefficients of selected principal components specific to image n
- B is the list of coefficient of non-selected principal components, common to all images

$$\tilde{\mathbf{x}}^{(n)} = \sum_{j=1}^{M} z_j^{(n)} \mathbf{u}_j + \sum_{j=M+1}^{D} b_j \mathbf{u}_j$$

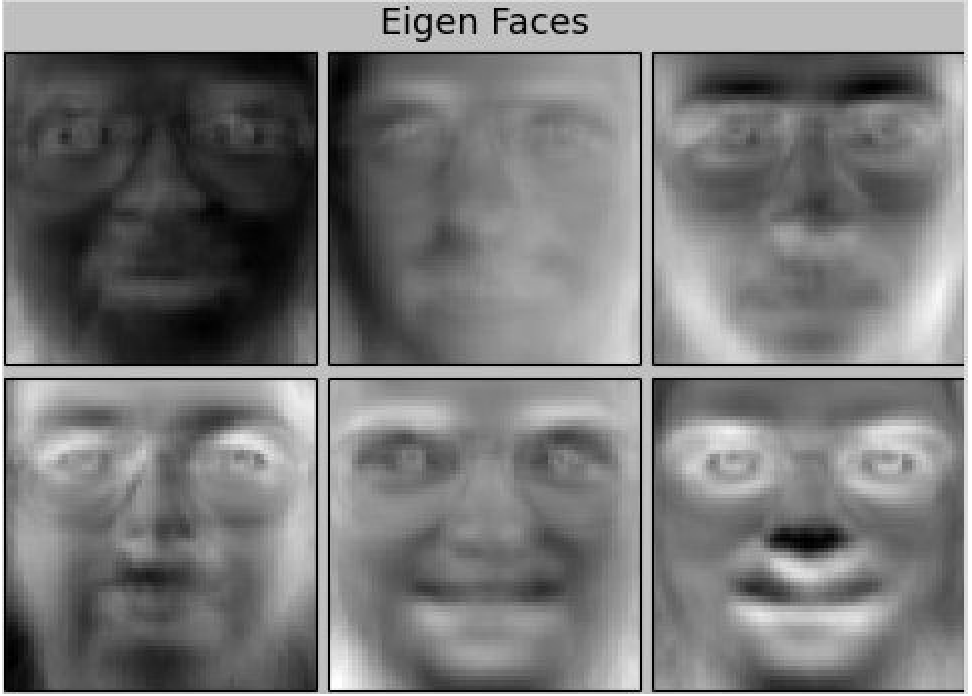
$$z_j^{(n)} = (\mathbf{x}^{(n)})^T \mathbf{u}_j; \quad b_j = \bar{\mathbf{x}}^T \mathbf{u}_j$$

Olivetti Faces Dataset - PCA

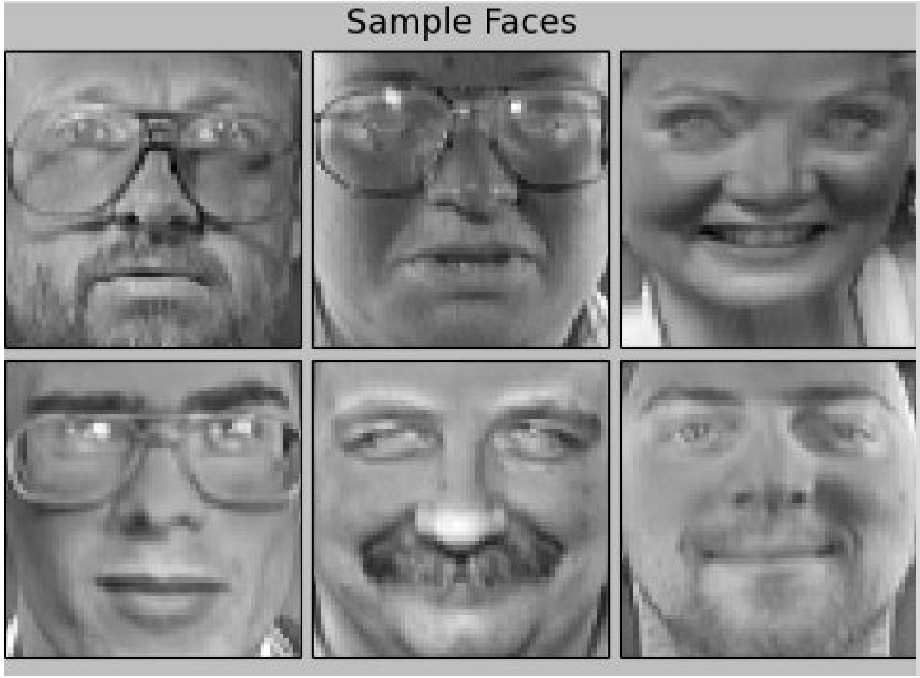
- We can now find the weights vector for faces in the dataset

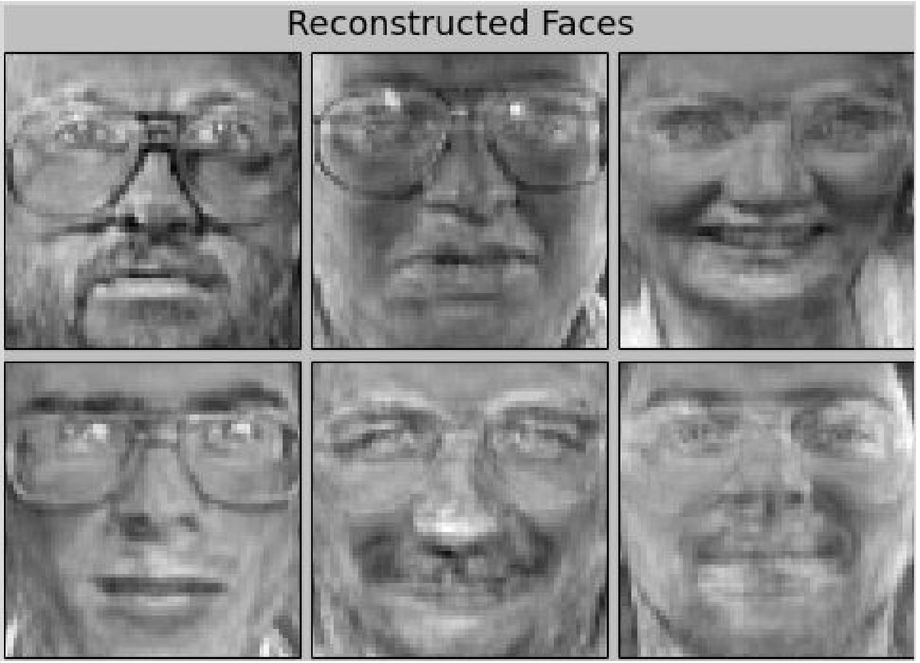
Original

Reconstructed with 200 components



Olivetti Faces Dataset - PCA





Original: dimension = 4096

New: dimension = 200

Bayesian Inference

- Basic concepts
  - Bayes’ theorem, bayesian modelling, conjugacy.
- Beta --- Binomial: conjugate prior
- Coin toss example
- Bayesian predictive distribution as ensembles

Bayes Theorem

- The **posterior** probability of θ, given our observation (x) is proportional to the **likelihood** times the **prior** probability of θ.

$$\boxed{P(\theta|x)=\frac{P(x|\theta)P(\theta)}{P(x)}}$$

Bayesian Modelling

$$P(\theta|\mathcal{D},m) = \frac{P(\mathcal{D}|\theta,m)P(\theta|m)}{P(\mathcal{D}|m)}$$

$P(\mathcal{D}|\theta,m)$	likelihood of parameters θ in model m
$P(\theta|m)$	prior probability of θ
$P(\theta|\mathcal{D},m)$	posterior of θ given data $\mathcal{D}$

Prediction:

$$P(x|\mathcal{D},m) = \int P(x|\theta,\mathcal{D},m)P(\theta|\mathcal{D},m)d\theta$$

Model Comparison:

$$P(m|\mathcal{D}) = \frac{P(\mathcal{D}|m)P(m)}{P(\mathcal{D})}$$

$$P(\mathcal{D}|m) = \int P(\mathcal{D}|\theta,m)P(\theta|m)\,d\theta$$

Conjugacy

If the posterior distributions $p(\theta|x)$ are in the same family as the prior probability distribution $p(\theta)$, the prior and posterior are then called **conjugate distributions**, and the prior is called a **conjugate prior** for the likelihood.

Binomial Data, Beta Prior

Suppose the prior distribution for θ is Beta(α1,α2) and the conditional distribution of X given θ is Bin(n, p). Then

$$P(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{(n-x)}$$

$$P(\theta) = \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)+\Gamma(\alpha_2)}\theta^{\alpha_1-1}(1-\theta)^{(\alpha_2-1)}$$

Binomial Data, Beta Prior (cont.)

We now calculate the posterior:

$$\text{posterior} \propto \text{likelihood} \times \text{prior}.$$

$$P(\theta|x) \propto P(x|\theta)P(\theta)$$

$$= \binom{n}{x} \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1) + \Gamma(\alpha_2)} \theta^{x+\alpha_1-1}(1-\theta)^{(n-x+\alpha_2-1)}$$

Binomial Data, Beta Prior (cont.)

Given x, $\binom{n}{x}\frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)+\Gamma(\alpha_2)}$ is a constant. Therefore,

$$P(\theta|x) \propto \theta^{x+\alpha_1-1}(1-\theta)^{(n-x+\alpha_2-1)}$$

We now recognize it as another Beta distribution with parameter (x+α1) and (n-x+α2): Beta(x+α1,n-x+α2).

Same family as the prior distribution: conjugate prior!

Coin toss example

- You have a coin that when flipped ends up head with probability θ and ends up tail with probability (1-θ).
- Trying to estimate θ, you flip the coin **14** times. It ends up head **10** times.
- What is the probability of: "*In the next two tosses we will get two heads in a row*"?
- Would you bet on "yes"?

Credit to: http://www.behind-the-enemy-lines.com/2008/01/are-you-bayesian-or-frequentist-or.html

Coin toss example (cont.) --- Frequentist approach

- Using frequentist statistics we would say that the best (maximum likelihood) estimate for θ is 10/14, i.e., θ≈0.714.
- In this case, the probability of two heads is 0.7142^2≈0.51 and it makes sense to bet for the event.
- Therefore, the frequentist will bet “yes”!

Coin toss example (Cont.) --- Bayesian

First let’s consider what likelihood function is. The coin toss follows a binomial distribution Bin(n,θ). Hence,

$$P(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{(n-x)}$$

In our case:

$$P(data|\theta) = \binom{14}{10} \theta^{10} (1-\theta)^4$$

Coin toss example (Cont.) --- Bayesian

As we have shown earlier, a **very convenient** prior distribution for binomial distribution is its **conjugate prior** : Beta distribution.

Let's put this prior on θ, with hyperparameter α1, α2:

$$P(\theta) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1) + \Gamma(\alpha_2)} \theta^{\alpha_1 - 1} (1 - \theta)^{(\alpha_2 - 1)}$$

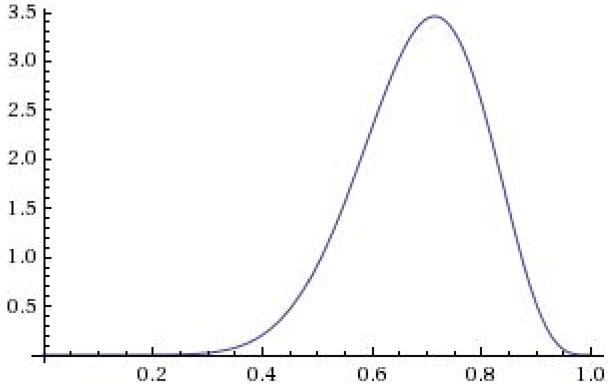
$$\Gamma(n) = (n-1)!$$

Coin toss example (Cont.) --- Bayesian

Therefore, the posterior distribution for θ is:

Beta(x+α1,n-x+α2) ---in our case --- Beta(10+α1,4+α2)

If we assume we know nothing about θ, then α1=α2=1. We plot the posterior distribution, i.e., (Beta(11,5)):



Coin toss example (Cont.) --- Two heads in a row

We perform prediction by integrating out our posterior belief on θ

$$Pr\{HH|data\} = \int_0^1 Pr\{HH|\theta\} \cdot P(\theta|data) d\theta$$

$$= \frac{1}{B(10+\alpha_1, 4+\alpha_2)} \int_0^1 \theta^2 \theta^{10+\alpha_1-1}(1-\theta)^{4+\alpha_2-1}$$

When $\alpha 1=\alpha 2=1$, this is 0.485. The Bayesian will bet “no”!

Coin toss example (Cont.) --- Model Comparison

Consider the following two models to fit the data:

M1 model using a fixed θ=0.5

M2 model employing a uniform prior over the unknown θ:

To choose which model is better, we need to compute the **marginal likelihood or model evidence**

$$Pr\{data|M\}$$

$$= \int_{\theta} Pr\{data|\theta, M\} Pr\{\theta|M\} d\theta$$

Coin toss example (Cont.) --- Model Comparison

Consider the following two models to fit the data:

M1 model using a fixed θ=0.5

M2 model employing a uniform prior over the unknown θ:

$$\underbrace{Pr\{data|M1\}}_{\text{marginal likelihood of M1}} = \binom{14}{10} 0.5^{10}(1-0.5)^4 \approx 0.0611$$

$$\underbrace{Pr\{data|M2\}}_{\text{marginal likelihood of M2}} = \int_0^1 \binom{14}{10} \theta^{10}(1-\theta)^4 d\theta$$

$$= \binom{14}{10} B(11,5) = \frac{14!}{10!4!}\frac{10!4!}{15!} \approx 0.066$$

slightly better model with an additional free parameter

Prediction as Ensemble

Given model M1 and M2, and their model evidence, we can do prediction in a form of model ensemble.

$$P(x|data) = \sum_i P(x|data, M_i)P(M_i|data)$$