Lecture 23
Bayesian regression methods

Bayes' rule; weight-based view and linear model

Prof. David A. Kofke
CE 500 – Modeling Potential-Energy Surfaces
Department of Chemical & Biological Engineering
University at Buffalo

University at Buffalo
The State University of New York

What color is the candy that I will select from this urn?

- It won't be white
- It is one of these colors

  black, red, green, blue, yellow, brown, orange, purple, gray
- The candy is not licorice
- The candy is M&Ms
- The color distribution of M&Ms is

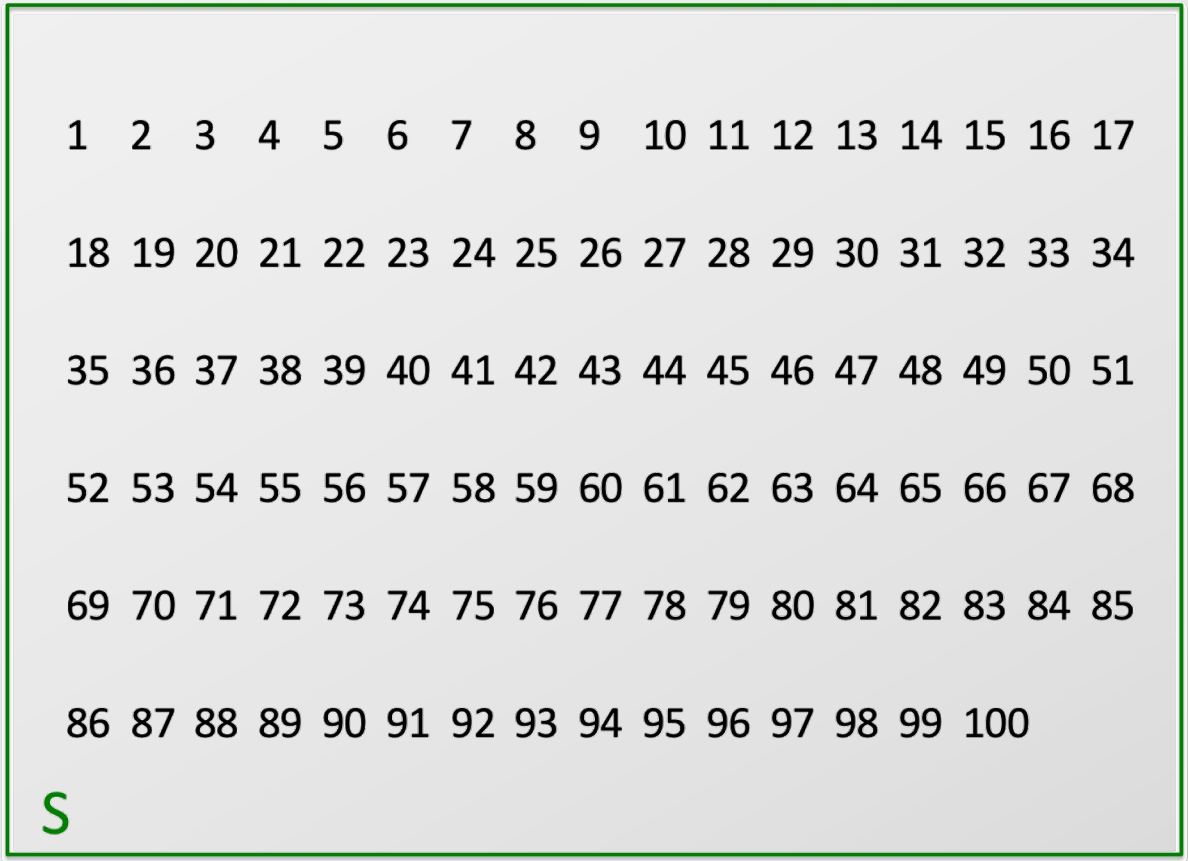
  24% blue, 20% orange, 16% green, 14% yellow, 13% red, and 13% brown
- What principle are you using to make your prediction?

Probability (like temperature) is a more subtle concept than you might realize

- Two interpretations are in common use
- Frequentist
  - Probability describes the number of times each outcome would occur if an infinite number of samples were taken on a population
  - It can change only by changing the population
- Bayesian
  - Probability quantifies how much information you have about an outcome
  - It can change as more information becomes available

Conditional probability is a core concept in the Bayesian framework

- Here is a space of outcomes of an event:
  - Select a number uniformly at random from 1 to 100

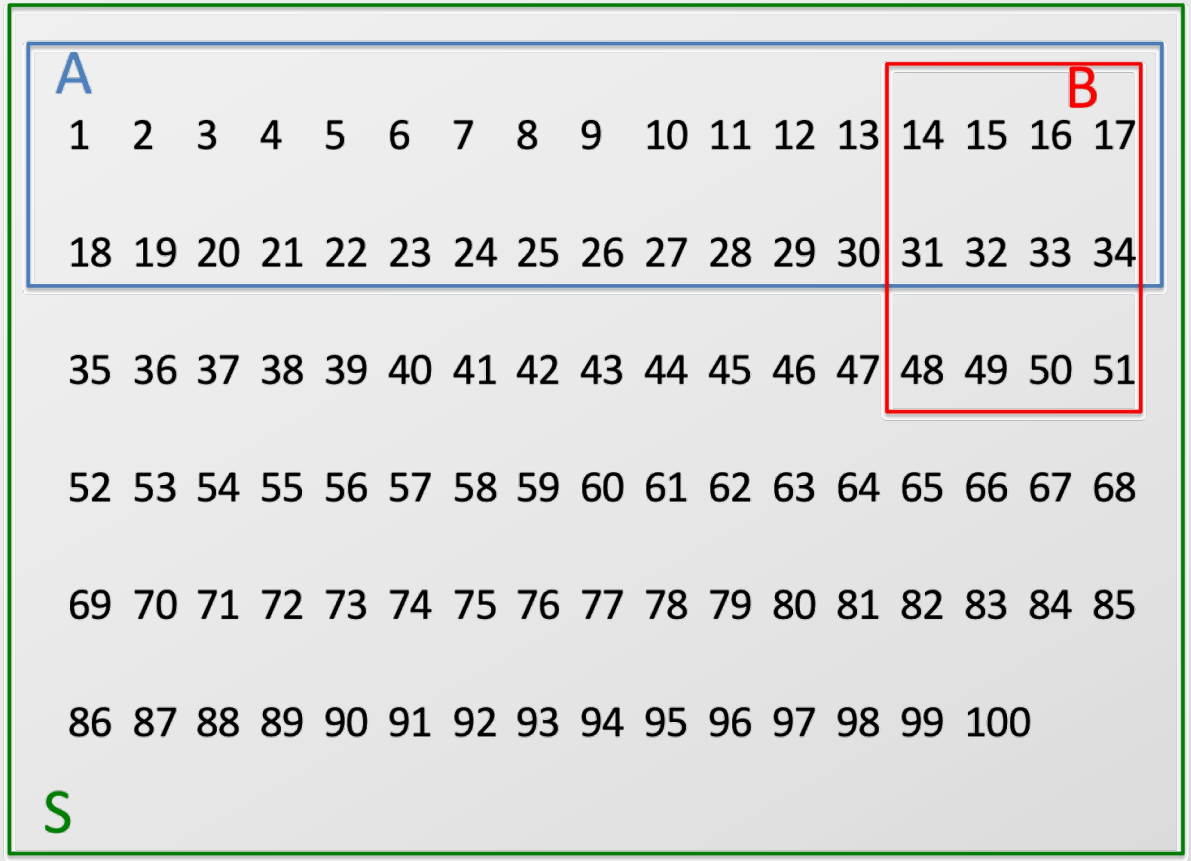


Conditional probability is a core concept in the Bayesian framework

- Events A and B occur when numbers are selected in the indicated regions, respectively
- What are

  P(A)?

  P(B)?

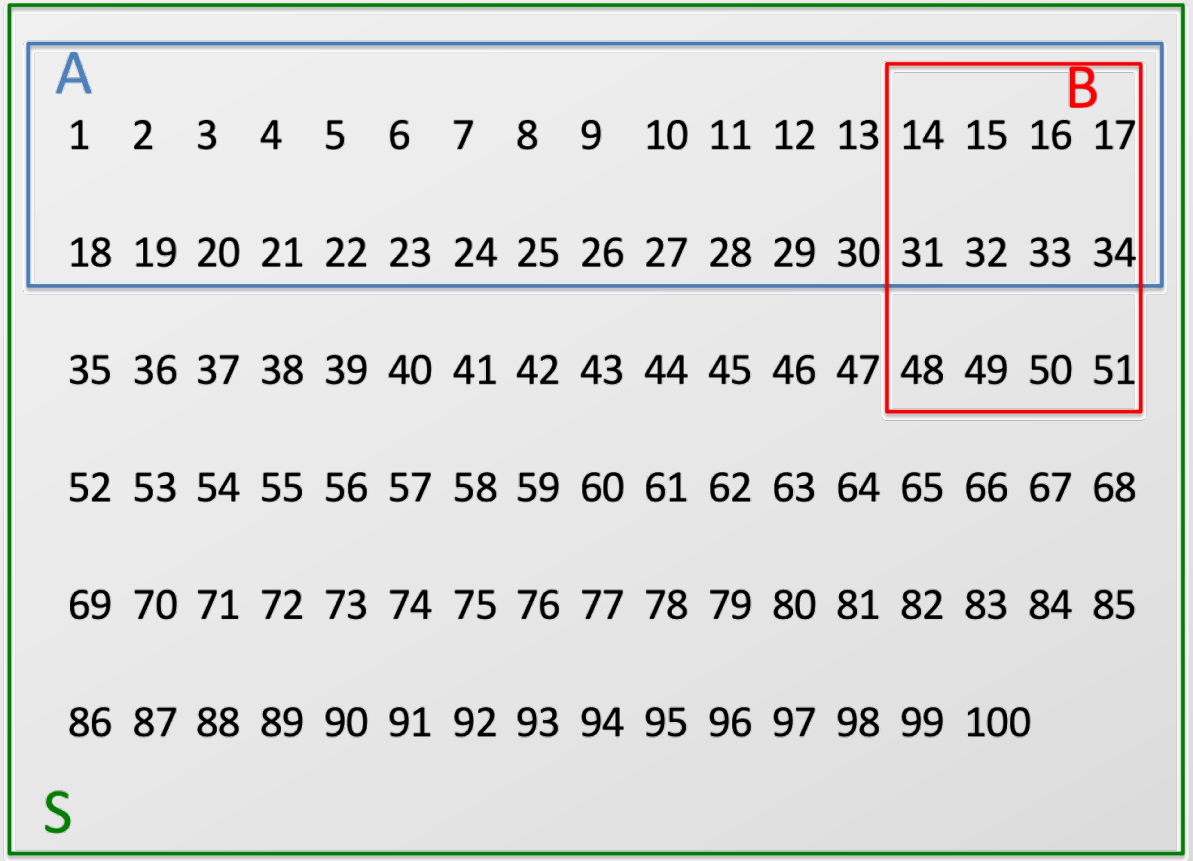


Conditional probability is a core concept in the Bayesian framework

- Events A and B occur when numbers are selected in the indicated regions, respectively
- What are

  $P(A) = 34/100 = 0.34$

  $P(B) = 12/100 = 0.12$

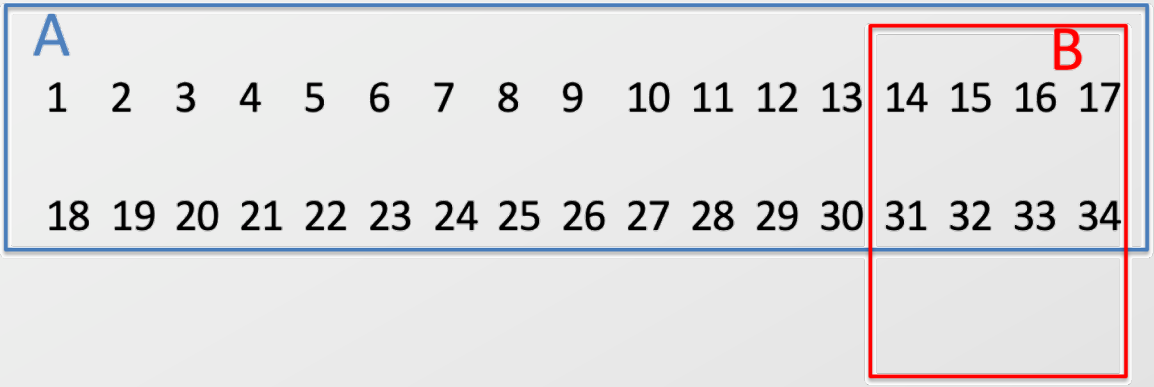


Conditional probability is a core concept in the Bayesian framework

- What is the probability of B, if we know that A occurred?

  Probability of B, *given* A

  P(B|A) = ?

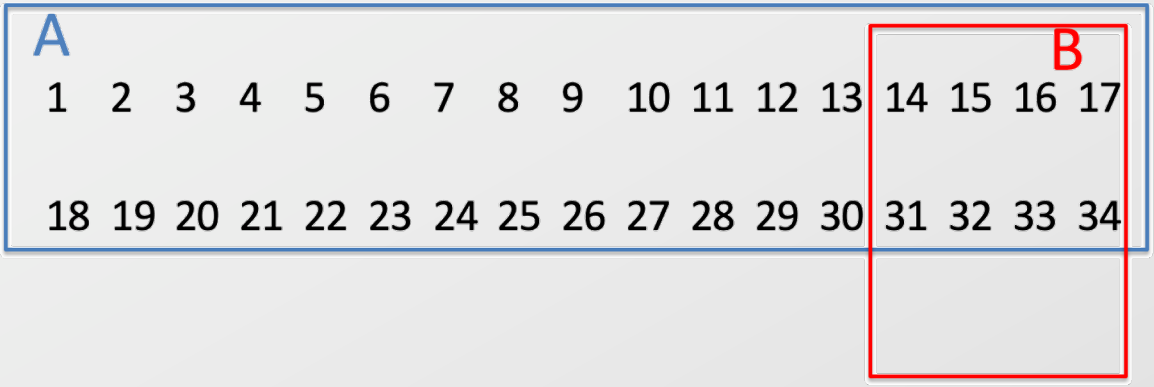


Conditional probability is a core concept in the Bayesian framework

- What is the probability of B, if we know that A occurred?

  Probability of B, *given* A

  P(B|A) = 8/34 = 0.24



Conditional probability is a core concept in the Bayesian framework

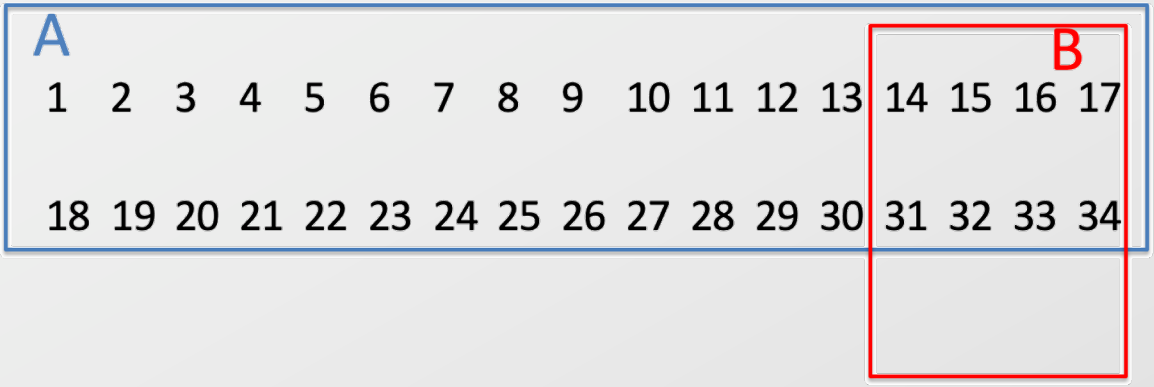
- What is the probability of B, if we know that A occurred?

  Probability of B, *given* A

  $P(B|A) = 8/34 = 0.24$

  $= P(A \cap B)/P(A)$

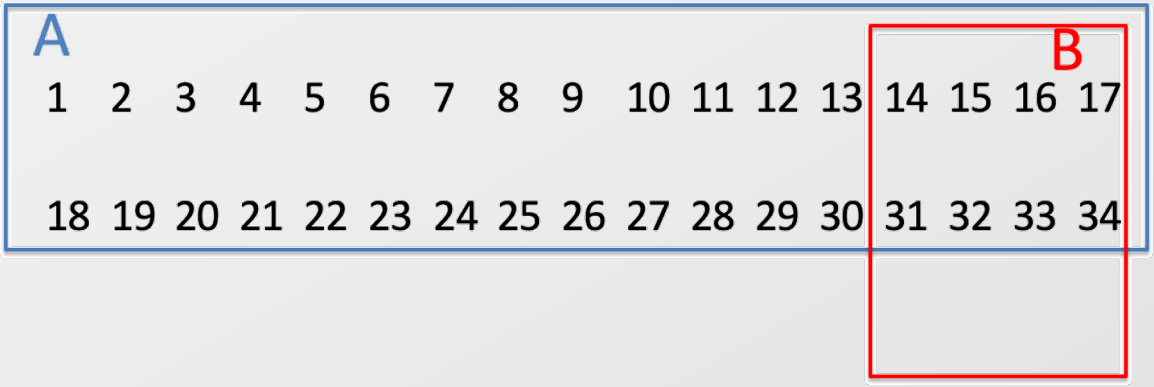
  $= (8/100)/(34/100)$



Conditional probability is a core concept in the Bayesian framework

- This is the general formula for the *conditional probability*

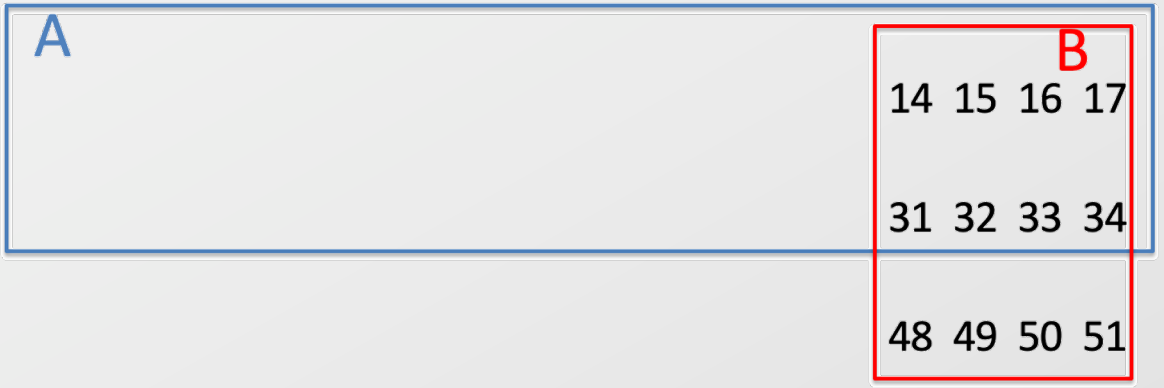
$$P(\mathrm{B}|\mathrm{A}) = \frac{P(\mathrm{B} \cap \mathrm{A})}{P(\mathrm{A})}$$



Conditional probability is a core concept in the Bayesian framework

- It works both ways
- What is

  P(A|B)?



Conditional probability is a core concept in the Bayesian framework

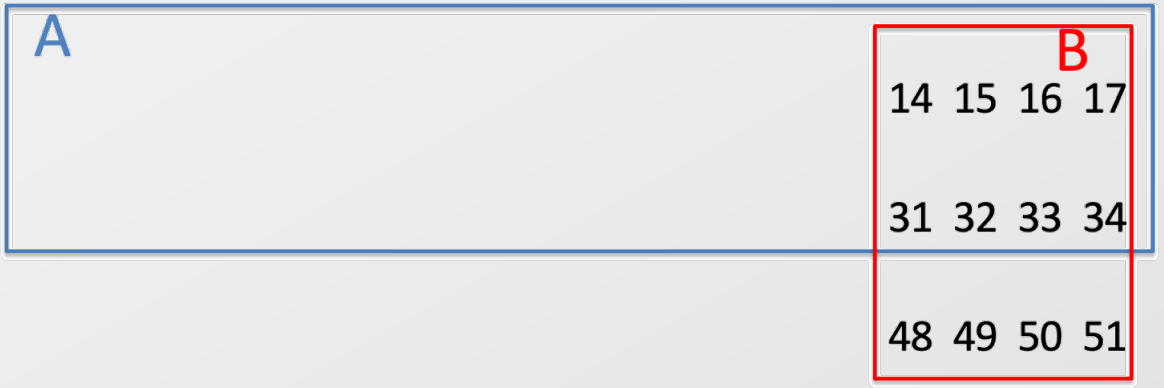
- It works both ways
- What is

  $P(A|B) = 8/12 = 0.67$

  $P(A \cap B)$?

  $P(B)$?

  $P(A \cap B)/P(B)$?



Conditional probability is a core concept in the Bayesian framework

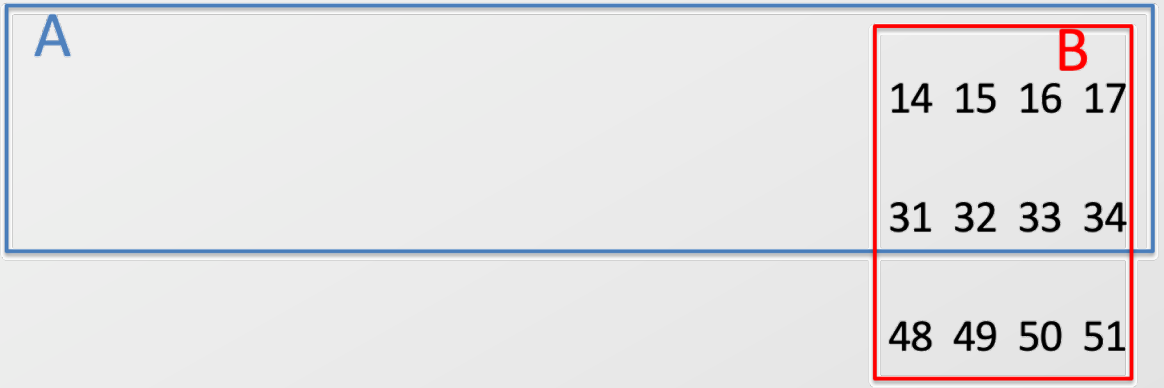
- It works both ways
- What is

$P(A|B) = 8/12 = 0.67$

$P(A \cap B) = 8/100$

$P(B) = 12/100$

$P(A \cap B)/P(B) = (8/100)/(12/100) = 8/12$



Conditional probability is a core concept in the Bayesian framework

- We have expressions for the two conditional probabilities

$$P(\mathrm{B}|\mathrm{A}) = \frac{P(\mathrm{B} \cap \mathrm{A})}{P(\mathrm{A})} \qquad P(\mathrm{A}|\mathrm{B}) = \frac{P(\mathrm{A} \cap \mathrm{B})}{P(\mathrm{B})}$$

- Or

$$P(\mathrm{B} \cap \mathrm{A}) = P(\mathrm{B}|\mathrm{A})P(\mathrm{A}) \qquad P(\mathrm{A} \cap \mathrm{B}) = P(\mathrm{A}|\mathrm{B})P(\mathrm{B})$$

- Of course, (A ∩ B) and (B ∩ A) are the same event, so

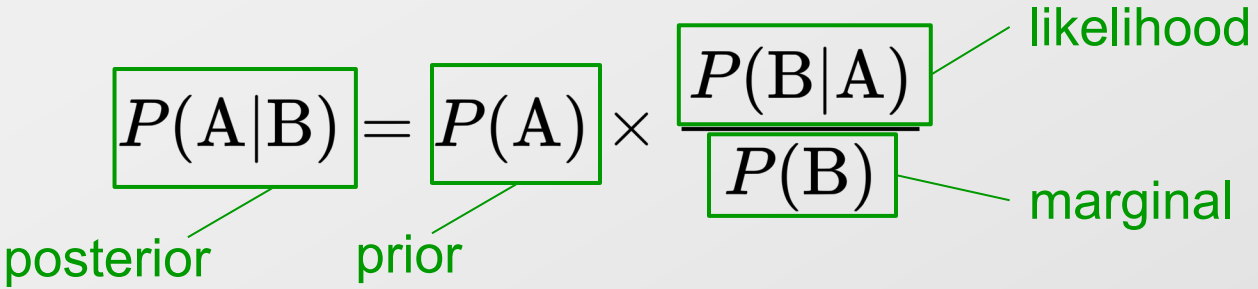
$$\boxed{P(\mathrm{B}|\mathrm{A})P(\mathrm{A}) = P(\mathrm{A}|\mathrm{B})P(\mathrm{B})}$$

  - This is *Bayes' rule*

Bayes' rule provides a framework to update probabilities when given new information

$$P(\mathrm{B}|\mathrm{A})P(\mathrm{A}) = P(\mathrm{A}|\mathrm{B})P(\mathrm{B})$$

- It's use is best understood when written in this form



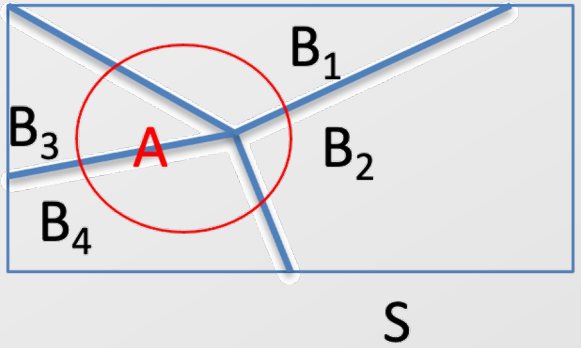
- The posterior is the updated probability in light of new information

The various parts of Bayes' formula can be explained in words

- In our applications, event A is a model parameter(s) **w** or outcome *y* having a particular value, and event B is data
- P(**w**) − this is the *prior probability*, the probability of the parameter value (or outcome *y*) in the absence of data
  - Could be a uniform distribution over an expected range
- P(**w**|data) − this is the *posterior probability*, informed by data
- P(data|**w**) − this is the *likelihood* that we'd observe the given data, given that the model parameter has the specified value
- P(data) − this is the *marginal probability* of the data, considering all possible parameter values

The marginal probability is usually obtained by integration, or normalization

$$P(\mathbf{w}|\text{data}) = P(\mathbf{w}) \times \frac{P(\text{data}|\mathbf{w})}{P(\text{data})}$$



- Consider all mutually exclusive B events

$$\begin{aligned} P(\mathrm{A}) &= P(\mathrm{A} \cap \mathrm{B}_1) + \cdots + P(\mathrm{A} \cap \mathrm{B}_n) \\ &= P(\mathrm{A}|\mathrm{B}_1)P(\mathrm{B}_1) + \cdots + P(\mathrm{A}|\mathrm{B}_n)P(\mathrm{B}_n) \end{aligned}$$

- Or more generally for our case

$$P(\text{data}) = \int P(\text{data}|\mathbf{w})P(\mathbf{w})d\theta$$

The marginal probability does not depend on **w**

Regression is a process of induction: specific → general

- Given a finite set of training data
- We want to infer a function or model $f(\mathbf{x})$ that provides accurate outputs for all possible input values
- The data are insufficient to do this perfectly, i.e., to solve for f
- Hence, we impose conditions, or assumptions, regarding the form of f
  - Otherwise, any function that goes through the points would be valid

Two approaches to restrict the model are in common use

- Restrict the class of functions that are considered
  - *e.g.*, linear functions of the variables in **x**
  - Requires a decision on the richness of the class of functions considered
    - Target may not be well described by the selected functions
    - Adding more functions risks overfitting
- Assign a *prior probability* to all trial functions
  - Higher probability assigned to more likely functions, e.g., those that are smooth versus those that are erratic
  - This admits and infinite range of functions; how to account for all of them? (Gaussian processes does the trick)

Rather than going straight to functions, a weight-space view can ease into the Bayesian approach

- Standard linear model

Input

$$\mathbf{x}^{(i)} = \begin{pmatrix} x_1^{(i)} \\ x_2^{(i)} \\ \vdots \\ x_p^{(i)} \end{pmatrix} \quad p$$

Design matrix

$$X = \begin{pmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_p^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots & x_p^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(n)} & x_2^{(n)} & \cdots & x_p^{(n)} \end{pmatrix}$$

$(\mathbf{x}^{(1)})^\top$ → first row

n {x,y} samples

p input values

Output

$$\mathbf{y} = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{pmatrix} \quad n$$

Weights

$$\mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_p \end{pmatrix} \quad p$$

- Gaussian noise is added to linear model to introduce deviations

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{w} \qquad y = f(\mathbf{x}) + \varepsilon \qquad \mathbf{y} = X\mathbf{w} + \boldsymbol{\varepsilon} \qquad \varepsilon \sim \mathcal{N}(0, \sigma_n^2)$$

Rasmussen & Williams define X as the transpose of the definition here

We start by writing the likelihood of the observed data, for the given inputs and weights

- Assuming independent data:

Deviation from model follows Gaussian pdf

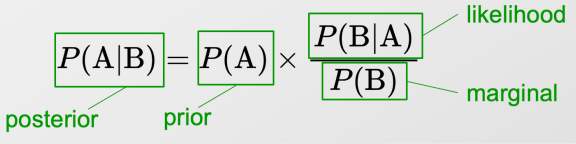
$$p(\mathbf{y} \mid X, \mathbf{w}) = \prod_{i=1}^{n} p(y_i \mid \mathbf{x}_i, \mathbf{w}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left(-\frac{\left(y_i - \mathbf{x}_i^\top \mathbf{w}\right)^2}{2\sigma_n^2}\right)$$
$$= \left(2\pi\sigma_n^2\right)^{-n/2} \exp\left(-\tfrac{1}{2\sigma_n^2}|\mathbf{y} - X\mathbf{w}|^2\right) = \mathcal{N}\left(X\mathbf{w}, \sigma_n^2 I\right)$$

- We need to specify a prior probability for the weights
  - Choose a zero-mean Gaussian with covariance matrix Σ_p

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \Sigma_p)$$

The Bayesian linear model specifies the weights via a posterior distribution, given the data

$$p(\mathbf{w} \mid \mathbf{y}, X) = p(\mathbf{w}) \frac{p(\mathbf{y} \mid X, \mathbf{w})}{p(\mathbf{y} \mid X)}$$



- The marginal probability is independent of $\mathbf{w}$
- Obtainable as normalization constant: $p(\mathbf{y} \mid X) = \int p(\mathbf{y} \mid X, \mathbf{w}) p(\mathbf{w}) d\mathbf{w}$

- The result can be expressed as a multivariate Gaussian

$$p(\mathbf{w} \mid \mathbf{y}, X) \sim \mathcal{N}(\bar{\mathbf{w}}, A^{-1})$$

- Mean: $\bar{\mathbf{w}} = \frac{1}{\sigma_n^2} A^{-1} X^\top \mathbf{y}$
- Covariance matrix: A^{-1}, $A = \sigma_n^{-2} X^\top X + \Sigma_{\mathrm{p}}^{-1}$

The recommended weights can be extracted in several ways

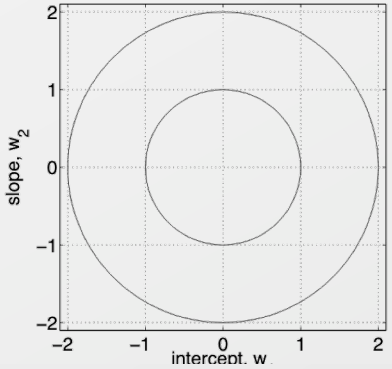
- Use the mean (expected value), or the mode (most likely value)
  - These are the same for the Normal distribution
- We obtain: $\bar{\mathbf{w}} = \left(X^\top X + \sigma_n^2 \Sigma_\mathrm{p}^{-1}\right)^{-1} X\mathbf{y}$
- Compare to result for ridge regression (Lecture 19)

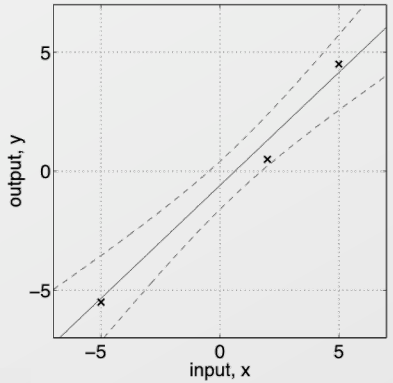
$$\hat{\mathbf{w}} = (X^\top X + \lambda I)^{-1} X^\top \mathbf{y}$$

- The prior on the weights imposes a regularization penalty!

Everything is given as a probability distribution

prior

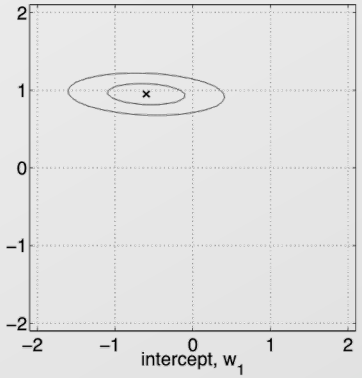




(b)

y(x)

posterior



(d)

Rasmussen & Williams

To make predictions in the Bayesian scheme, we can average over the entire distribution of weights

- This again results in a Gaussian for the predicted values

$$p(f_* \mid \mathbf{x}_*, X, \mathbf{y}) = \int p(f_* \mid \mathbf{x}_*, \mathbf{w}) p(\mathbf{w} \mid X, \mathbf{y}) d\mathbf{w}$$

$$= \mathcal{N}(\underbrace{\sigma_n^{-2} \mathbf{x}_*^\top A^{-1} X^\top \mathbf{y}}_{\text{mean}}, \underbrace{\mathbf{x}_*^\top A^{-1} \mathbf{x}_*}_{\text{variance}})$$

predicted value — input vector

- For an infinite-variance prior ($\Sigma_p^{-1} \rightarrow 0$)

$$A = \sigma_n^{-2} X^\top X + \Sigma_p^{-1}$$

$$\sigma_n^{-2} A^{-1} = (X^\top X)^{-1}$$

and the mean $f_*(\mathbf{x}_*)$ is $\mathbf{x}_*^\top (X^\top X)^{-1} X^\top \mathbf{y} = \mathbf{x}_*^\top \bar{\mathbf{w}}$

Features (a.k.a. basis functions) are a way to add versatility to the model formulation

- Expand the p-dimensional $\mathbf{x}$ vector into an N-dimensional space of variables $\boldsymbol{\phi}(\mathbf{x})$
- E.g., instead of a straight-line fit to a single x variable, use a polynomial fit

$$\mathbf{x} = (x) \qquad \boldsymbol{\phi}(\mathbf{x}) = \begin{pmatrix} 1 \\ x \\ x^2 \\ \vdots \\ x^N \end{pmatrix} \qquad f(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{w}$$

vector of length N

Analysis using features ϕ proceeds exactly as for case working directly with input vector x

- Replace X with feature matrix

$$\Phi(X) = \begin{pmatrix} \phi_1^{(1)}(\mathbf{x}) & \cdots & \phi_N^{(1)}(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ \phi_1^{(n)}(\mathbf{x}) & \cdots & \phi_N^{(n)}(\mathbf{x}) \end{pmatrix}$$

$n \times N$ matrix

- The predictive distribution is

$$f_* \mid \mathbf{x}_*, X, \mathbf{y} \sim N(\sigma_n^{-2}\boldsymbol{\phi}_*^\top A^{-1}\Phi\mathbf{y}, \boldsymbol{\phi}_*^\top A^{-1}\boldsymbol{\phi}_*)$$

$$\boldsymbol{\phi}_* \equiv \boldsymbol{\phi}(\mathbf{x}_*) \qquad A = \sigma_n^{-2}\Phi^\top\Phi + \Sigma_\mathrm{p}^{-1}$$

The result may be reformulated to introduce a *kernel* in lieu of the features

- Inversion of $N \times N$ matrix A may be inconvenient
- The distribution can be rewritten as

$$f_* \mid \mathbf{x}_*, X, \mathbf{y} \sim \mathcal{N}\left(\boldsymbol{\phi}_*^\top \Sigma_p \Phi^\top \left(K + \sigma_n^2 I\right)^{-1} \mathbf{y}\right.$$
$$\left. \boldsymbol{\phi}_*^\top \Sigma_p \boldsymbol{\phi}_* - \boldsymbol{\phi}_*^\top \Sigma_p \Phi^\top \left(K + \sigma_n^2 I\right)^{-1} \Phi \Sigma_p \boldsymbol{\phi}_*\right)$$

  - Where the kernel matrix is defined $K = \Phi \Sigma_p \Phi^\top$
  - This instead requires inversion of an $n \times n$ matrix
  - Desirable if feature space is larger than amount of data $N > n$

The result may be reformulated to introduce a *kernel* in lieu of the features

- Inversion of $N \times N$ matrix A may be inconvenient
- The distribution can be rewritten as

$$f_* \mid \mathbf{x}_*, X, \mathbf{y} \sim \mathcal{N}\Big(\boldsymbol{\phi}_*^\top \Sigma_p \Phi^\top \big(K + \sigma_n^2 I\big)^{-1} \mathbf{y}$$
$$\boldsymbol{\phi}_*^\top \Sigma_p \boldsymbol{\phi}_* - \boldsymbol{\phi}_*^\top \Sigma_p \Phi^\top \big(K + \sigma_n^2 I\big)^{-1} \Phi \Sigma_p \boldsymbol{\phi}_*\Big)$$

- Note that the features always enter in a specific way
- Define covariance function, or *kernel* $k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x}) \Sigma_p \boldsymbol{\phi}(\mathbf{x}')$
- Express as a dot product: $k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\psi}(\mathbf{x}) \cdot \boldsymbol{\psi}(\mathbf{x}'), \quad \boldsymbol{\psi}(\mathbf{x}) \equiv \Sigma_p^{1/2} \boldsymbol{\phi}(\mathbf{x})$

The *kernel trick* replaces inner products in x with the kernel operation, *k*(x,x')

- This allows model to bypass feature vectors entirely in lieu of the kernel
- The kernel then becomes the central focus of model development

Suggested Reading/Viewing

- C. E. Rasmussen & C. K. I. Williams, *Gaussian Processes for Machine Learning*, the MIT Press, 2006. Chapters 1 and 2.
  - www.GaussianProcess.org/gpml
- *Quantum Chemistry in the Age of Machine Learning*, edited by P. O. Dral
  - *Chapter 9. Kernel Methods,* Max Pinheiro Jr. and Pavlo O. Dral
  - *Chapter 10. Bayesian Inference*, Wei Liang and Hongsheng Dai
  - Posted on UBLearns