Machine Learning

Math Essentials

Areas of math essential to machine learning

- Machine learning is part of both ***statistics*** and computer science
  - Probability
  - Statistical inference
  - Validation
  - Estimates of error, confidence intervals
- ***Linear algebra***
  - Hugely useful for compact representation of linear transformations on data
  - Dimensionality reduction techniques
- ***Optimization*** theory

Why worry about the math?

- There are lots of easy-to-use machine learning packages out there.
- After this course, you will know how to apply several of the most general-purpose algorithms.

HOWEVER

- To get really useful results, you need good mathematical intuitions about certain general machine learning principles, as well as the inner workings of the individual algorithms.

Why worry about the math?

These intuitions will allow you to:

- Choose the right algorithm(s) for the problem
- Make good choices on parameter settings, validation strategies
- Recognize over- or underfitting
- Troubleshoot poor / ambiguous results
- Put appropriate bounds of confidence / uncertainty on results
- Do a better job of coding algorithms or incorporating them into more complex analysis pipelines

Notation

$a \in A$	*set membership*: *a* is *member* of set *A*
$\| B \|$	*cardinality*: number of items in set *B*
$\|\| \mathbf{v} \|\|$	*norm*: length of vector *v*
$\sum$	*summation*
$\int$	*integral*
$\Re$	the set of *real* numbers

$\Re^n$	*real number space* of dimension *n*
	n = 2 : plane or 2-space
	n = 3 : 3- (dimensional) space
	n > 3 : *n*-space or *hyperspace*

Notation

- $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$, $\mathbf{u}$, $\mathbf{v}$ *vector* (bold, lower case)
- $\mathbf{A}$, $\mathbf{B}$, $\mathbf{X}$ *matrix* (bold, upper case)
- $y = f(x)$ *function* (*map*): assigns unique value in range of y to each value in domain of x
- dy / dx *derivative* of y with respect to single variable x
- $y = f(\mathbf{x})$ *function* on multiple variables, i.e. a vector of variables; *function* in n-space
- $\partial y / \partial x_i$ *partial derivative* of y with respect to element i of vector $\mathbf{x}$

The concept of probability

Intuition:

- In some process, several outcomes are possible. When the process is repeated a large number of times, each outcome occurs with a characteristic *relative frequency*, or *probability*. If a particular outcome happens more often than another outcome, we say it is more probable.

The concept of probability

Arises in two contexts:

- In actual repeated experiments.
  - Example: You record the color of 1000 cars driving by. 57 of them are green. You *estimate* the probability of a car being green as 57 / 1000 = 0.0057.
- In idealized conceptions of a repeated process.
  - Example: You consider the behavior of an unbiased six-sided die. The *expected* probability of rolling a 5 is 1 / 6 = 0.1667.
  - Example: You need a model for how people's heights are distributed. You choose a normal distribution (bell-shaped curve) to represent the *expected* relative probabilities.

Probability spaces

A *probability space* is a *random process* or *experiment* with three components:

- Ω, the set of possible *outcomes* O
  - ◆ number of possible outcomes = $| \Omega | = N$
- F, the set of possible *events* E
  - ◆ an event comprises 0 to N outcomes
  - ◆ number of possible events = $| F | = 2^N$
- P, the *probability distribution*
  - ◆ function mapping each outcome and event to real number between 0 and 1 (the *probability* of O or E)
  - ◆ probability of an event is *sum* of probabilities of possible outcomes in event

Axioms of probability

1. Non-negativity:

   for any event $E \in F$, $p(E) \geq 0$

2. All possible outcomes:

   $p(\Omega) = 1$

3. Additivity of disjoint events:

   for all events $E, E' \in F$ where $E \cap E' = \varnothing$,
   $p(E \cup E') = p(E) + p(E')$

Types of probability spaces

Define | Ω | = number of possible outcomes

- Discrete space | Ω | is finite
  - Analysis involves *summations* ($\sum$)
- Continuous space | Ω | is infinite
  - Analysis involves *integrals* ($\int$)

Example of discrete probability space

Single roll of a six-sided die

- 6 possible outcomes: $O = 1, 2, 3, 4, 5$, or 6
- $2^6 = 64$ possible events
  - example: $E = (O \in \{ 1, 3, 5 \})$, i.e. outcome is odd
- If die is fair, then probabilities of outcomes are equal

$$p(1) = p(2) = p(3) =$$
$$p(4) = p(5) = p(6) = 1 / 6$$

  - example: probability of event E = (outcome is odd) is
    $p(1) + p(3) + p(5) = 1 / 2$

Example of discrete probability space

Three consecutive flips of a coin

- 8 possible outcomes: O = HHH, HHT, HTH, HTT, THH, THT, TTH, TTT
- 2^8 = 256 possible events
  - ◆ example: $E = (O \in \{ \text{HHT, HTH, THH} \})$, i.e. exactly two flips are heads
  - ◆ example: $E = (O \in \{ \text{THT, TTT} \})$, i.e. the first and third flips are tails
- If coin is fair, then probabilities of outcomes are equal

  p(HHH) = p(HHT) = p(HTH) = p(HTT) =
  p(THH) = p(THT) = p(TTH) = p(TTT) = 1 / 8

  - ◆ example: probability of event E = (exactly two heads) is

    p(HHT) + p(HTH) + p(THH) = 3 / 8

Example of continuous probability space

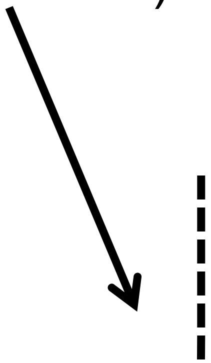
Height of a randomly chosen American male

- Infinite number of possible outcomes: *O* has some single value in range 2 feet to 8 feet
- Infinite number of possible events
  - ◆ example: $E = (O \mid O < 5.5 \text{ feet})$, i.e. individual chosen is less than 5.5 feet tall
- Probabilities of outcomes are not equal, and are described by a continuous function, $p(O)$

Example of continuous probability space

Height of a randomly chosen American male

- Probabilities of outcomes O are not equal, and are described by a continuous function, $p(O)$
- $p(O)$ is a *relative*, not an *absolute* probability
  - $p(O)$ for any particular O is zero
  - $\int p(O)$ from $O = -\infty$ to ∞ (i.e. area under curve) is 1
  - example: $p(O = 5'8'') > p(O = 6'2'')$
  - example: $p(O < 5'6'') = (\int p(O)$ from $O = -\infty$ to $5'6'') \approx 0.25$



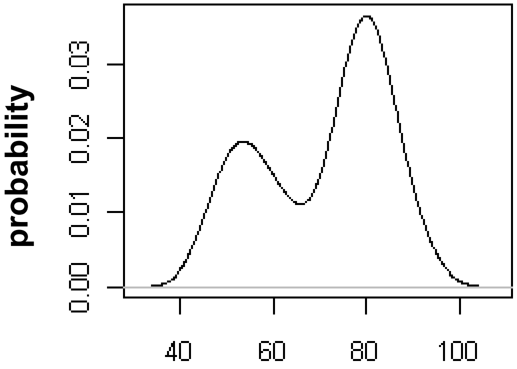
Probability distributions

- Discrete: *probability mass function* (pmf)

  example:
  sum of two
  fair dice

- Continuous: *probability density function* (pdf)

  example:
  waiting time between
  eruptions of Old Faithful
  (minutes)



Random variables

- A random variable X is a function that associates a number x with each outcome O of a process
  - Common notation: $X(O) = x$, or just $X = x$
- Basically a way to redefine (usually simplify) a probability space to a new probability space
  - X must obey axioms of probability (over the possible values of x)
  - X can be discrete or continuous
- Example: X = number of heads in three flips of a coin
  - Possible values of X are 0, 1, 2, 3
  - $p(X = 0) = p(X = 3) = 1 / 8$ $p(X = 1) = p(X = 2) = 3 / 8$
  - Size of space (number of "outcomes") reduced from 8 to 4
- Example: X = average height of five randomly chosen American men
  - Size of space unchanged (X can range from 2 feet to 8 feet), but pdf of X different than for single man

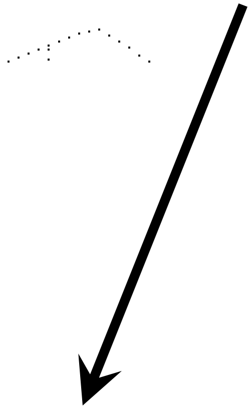
Multivariate probability distributions

- Scenario
  - Several random processes occur (doesn’t matter whether in parallel or in sequence)
  - Want to know probabilities for each possible combination of outcomes
- Can describe as *joint probability* of several random variables
  - Example: two processes whose outcomes are represented by random variables *X* and *Y*. Probability that process *X* has outcome *x* and process *Y* has outcome *y* is denoted as:

$$p(X = x, Y = y)$$

Example of multivariate distribution

joint probability: *p*(*X* = minivan, *Y* = European) = 0.1481



Multivariate probability distributions

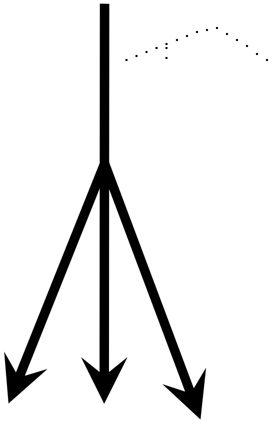
- *Marginal* probability
  - Probability distribution of a single variable in a joint distribution
  - Example: two random variables X and Y:

    $p(\ X = x\ ) = \sum_{b=\text{all values of } Y} p(\ X = x,\ Y = b\ )$
- *Conditional* probability
  - Probability distribution of one variable *given* that another variable takes a certain value
  - Example: two random variables X and Y:

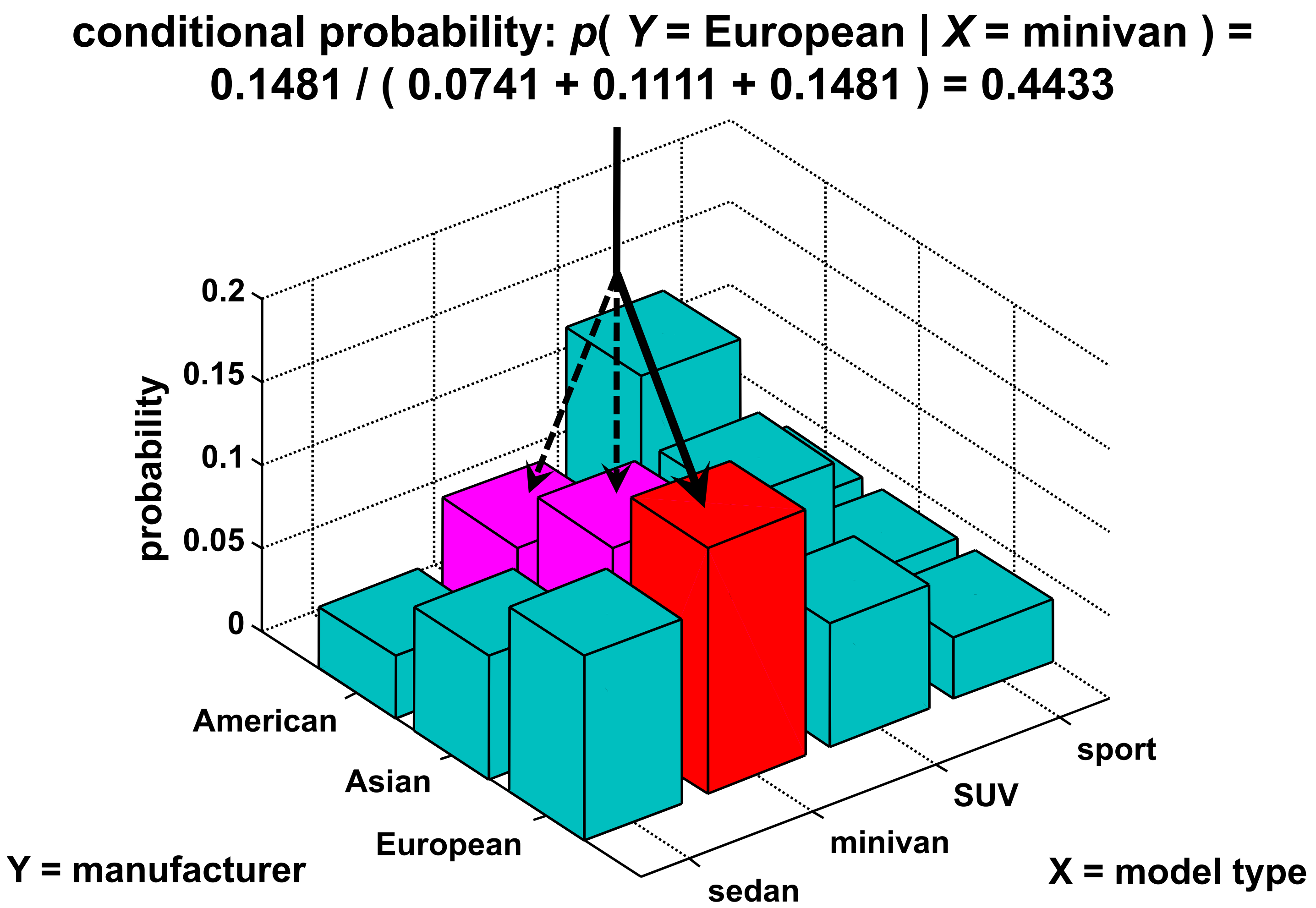
    $p(\ X = x \mid Y = y\ ) = p(\ X = x,\ Y = y\ ) \,/\, p(\ Y = y\ )$

Example of marginal probability

marginal probability: $p($ X = minivan $)$ = 0.0741 + 0.1111 + 0.1481 = 0.3333

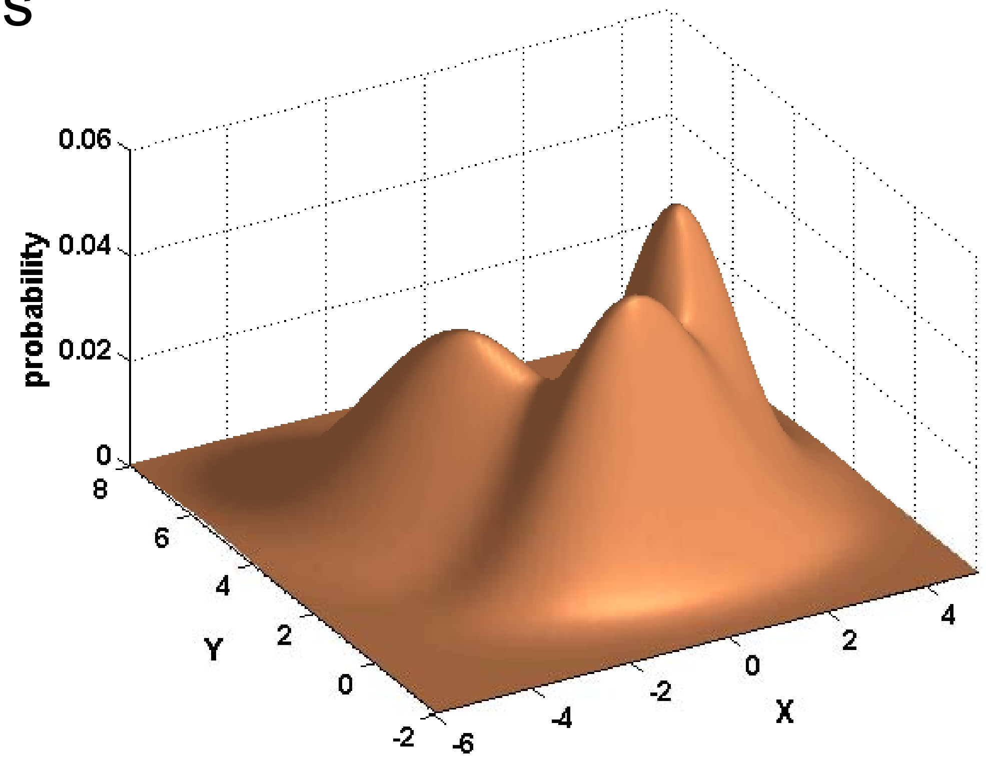


Example of conditional probability


conditional probability: $p(\ Y = \text{European} \mid X = \text{minivan}\ ) =$
$0.1481 / (\ 0.0741 + 0.1111 + 0.1481\ ) = 0.4433$

Continuous multivariate distribution

- Same concepts of joint, marginal, and conditional probabilities apply (except use integrals)
- Example: three-component Gaussian mixture in two dimensions



Expected value

Given:

- A discrete random variable X, with possible values $x = x_1, x_2, \ldots x_n$
- Probabilities $p(X = x_i)$ that X takes on the various values of x_i
- A function $y_i = f(x_i)$ defined on X

The *expected value* of f is the probability-weighted "average" value of $f(x_i)$:

$$E(f) = \sum_i p(x_i) \cdot f(x_i)$$

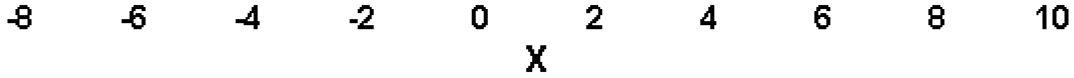
Example of expected value

- Process: game where one card is drawn from the deck
  - If face card, dealer pays you $10
  - If not a face card, you pay dealer $4
- Random variable X = { face card, not face card }
  - p(face card) = 3/13
  - p(not face card) = 10/13
- Function $f(X)$ is payout to you
  - f(face card) = 10
  - f(not face card) = -4
- *Expected value* of payout is:

$$E(f) = \sum_i p(x_i) \cdot f(x_i) = 3/13 \cdot 10 + 10/13 \cdot -4 = -0.77$$

Expected value in continuous spaces

$$\mathrm{E}(f) = \int_{x\,=\,a\,\rightarrow\,b} p(x) \cdot f(x)$$



Common forms of expected value (1)

- Mean (μ)

  $f(x_i) = x_i \;\Rightarrow\; \mu = \mathrm{E}(f) = \sum_i p(x_i) \cdot x_i$

  – Average value of $X = x_i$, taking into account probability of the various x_i

  – Most common measure of "center" of a distribution

- Compare to formula for mean of an actual sample

$$\mu = \frac{1}{N} \sum_{i=1}^{n} x_i$$

Common forms of expected value (2)

- Variance (σ^2)

  $f(x_i) = (x_i - \mu) \quad \Rightarrow \qquad \sigma^2 = \sum_i p(x_i) \cdot (x_i - \mu)^2$

  - Average value of squared deviation of $X = x_i$ from mean μ, taking into account probability of the various x_i
  - Most common measure of “spread” of a distribution
  - σ is the *standard deviation*

- Compare to formula for variance of an actual sample

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

Common forms of expected value (3)

- Covariance

  $f(x_i) = (x_i - \mu_x), \quad g(y_i) = (y_i - \mu_y) \quad \Rightarrow$

  $\text{cov}(x, y) = \sum_i p(x_i , y_i) \cdot (x_i - \mu_x) \cdot (y_i - \mu_y)$

  – Measures tendency for *x* and *y* to deviate from their means in same (or opposite) directions at same time

no covariance

high (positive) covariance

- Compare to formula for covariance of actual samples

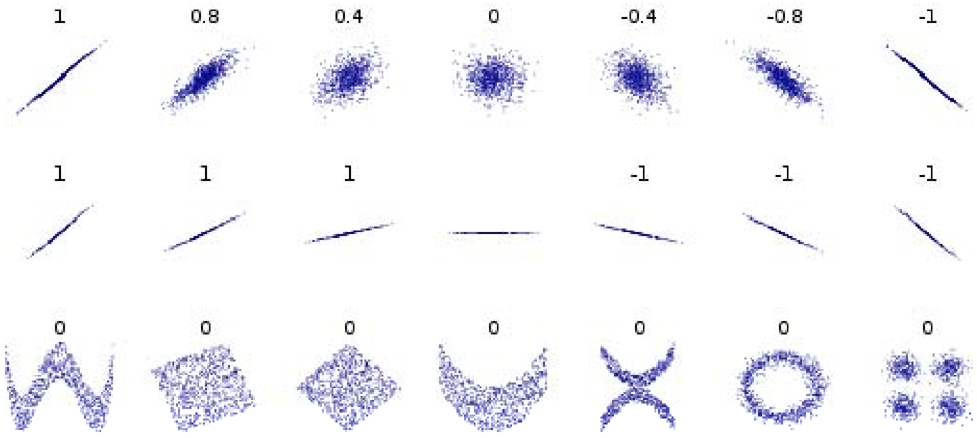
$$\text{cov}(x, y) = \frac{1}{N-1} \sum_{i=1}^{n} (x_i - \mu_x)(y_i - \mu_y)$$

Correlation

- Pearson's correlation coefficient is covariance normalized by the standard deviations of the two variables

$$\mathrm{corr}(x, y) = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y}$$

  - Always lies in range -1 to 1
  - Only reflects *linear dependence* between variables



Linear dependence with noise

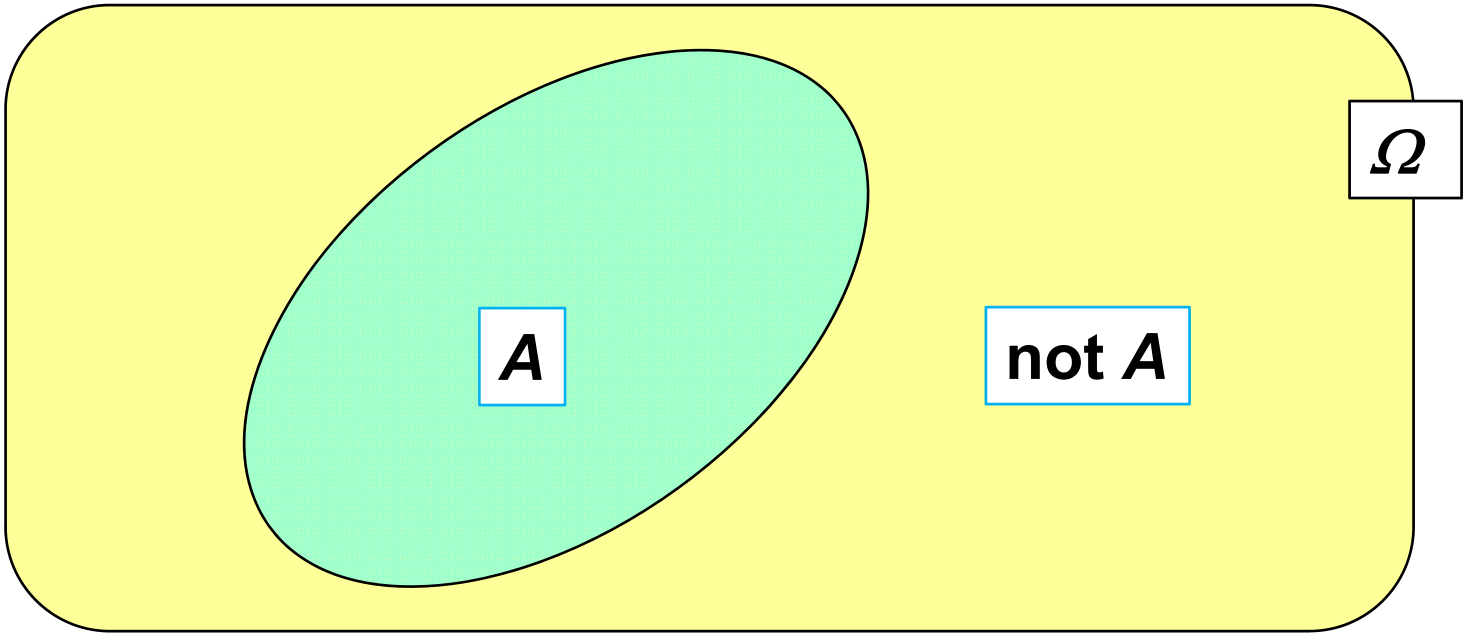
Linear dependence without noise

Various nonlinear dependencies

Complement rule

Given: event *A*, which can occur or not

$$p(\text{ not } A\ ) = 1 - p(\ A\ )$$



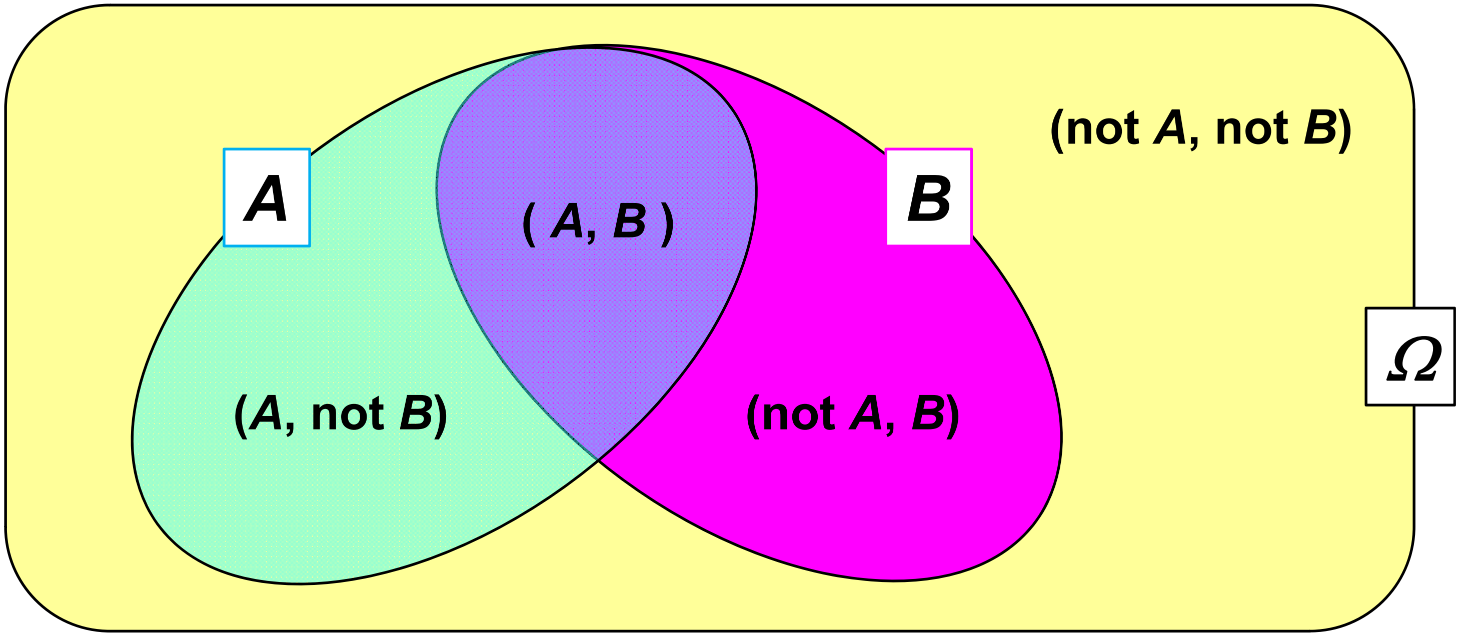
areas represent relative probabilities

Product rule

Given: events *A* and *B*, which can co-occur (or not)

$$p(A, B) = p(A \mid B) \cdot p(B)$$

(same expression given previously to define conditional probability)



areas represent relative probabilities

Example of product rule

- Probability that a man has white hair (event A) and is over 65 (event B)
  - $p(\ B\ ) = 0.18$
  - $p(\ A \mid B\ ) = 0.78$
  - $p(\ A, B\ ) = p(\ A \mid B\ ) \cdot p(\ B\ ) =$

    $0.78 \cdot 0.18 =$

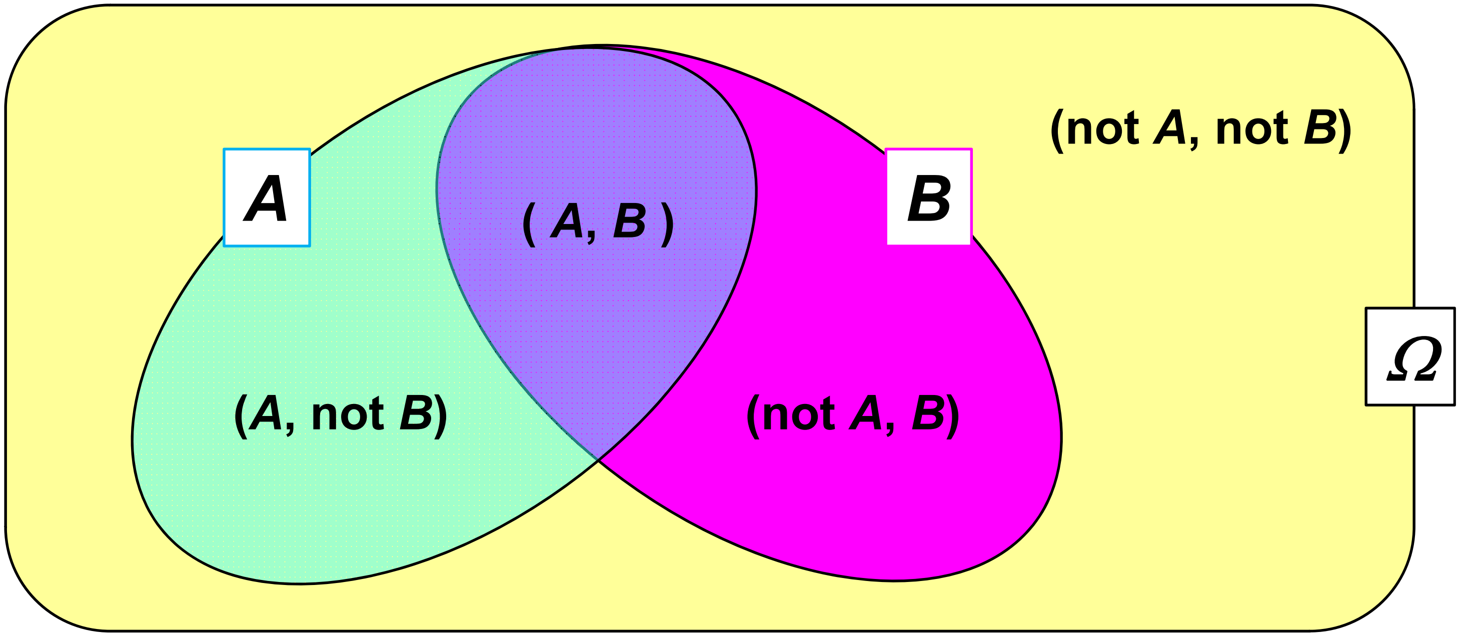
    0.14

Rule of total probability

Given: events *A* and *B*, which can co-occur (or not)

$$p(\ A\ ) = p(\ A, B\ ) + p(\ A, \text{not } B\ )$$

(same expression given previously to define marginal probability)

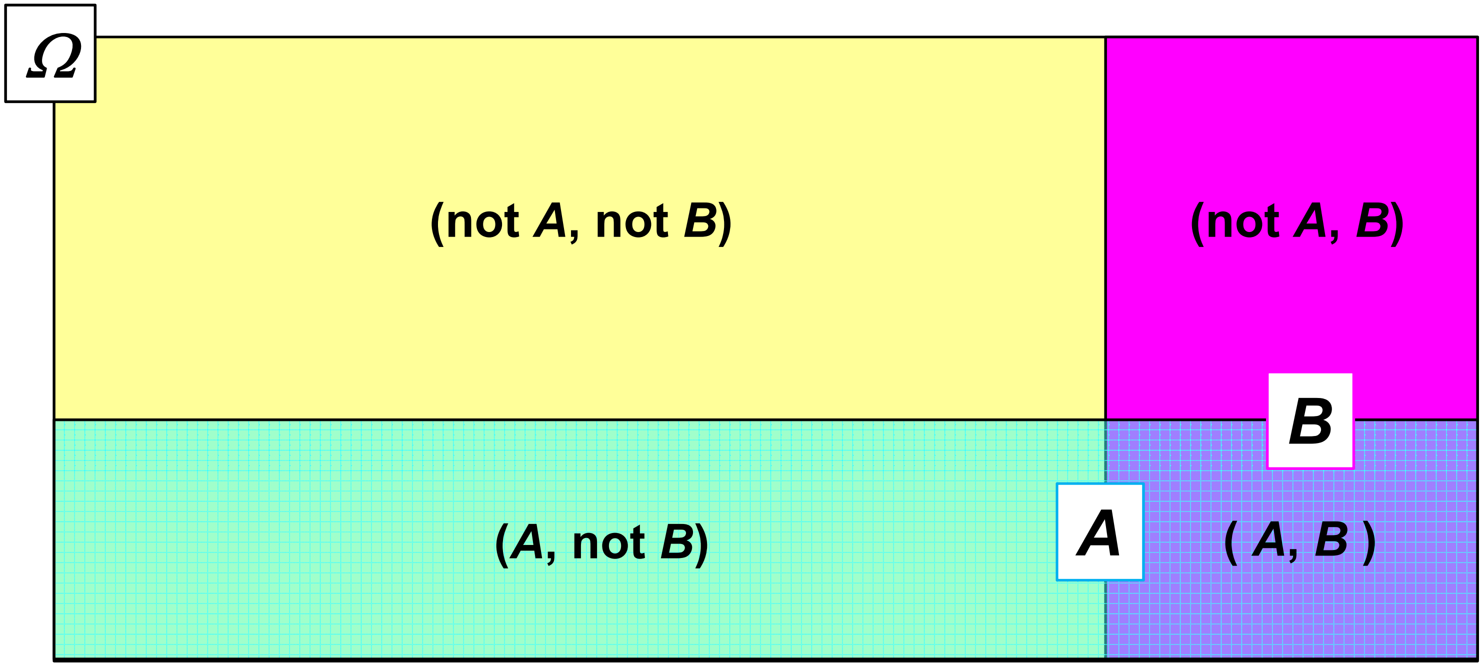


areas represent relative probabilities

Independence

Given: events A and B, which can co-occur (or not)

$$p(A \mid B) = p(A) \quad \text{or} \quad p(A, B) = p(A) \cdot p(B)$$



areas represent relative probabilities

Examples of independence / dependence

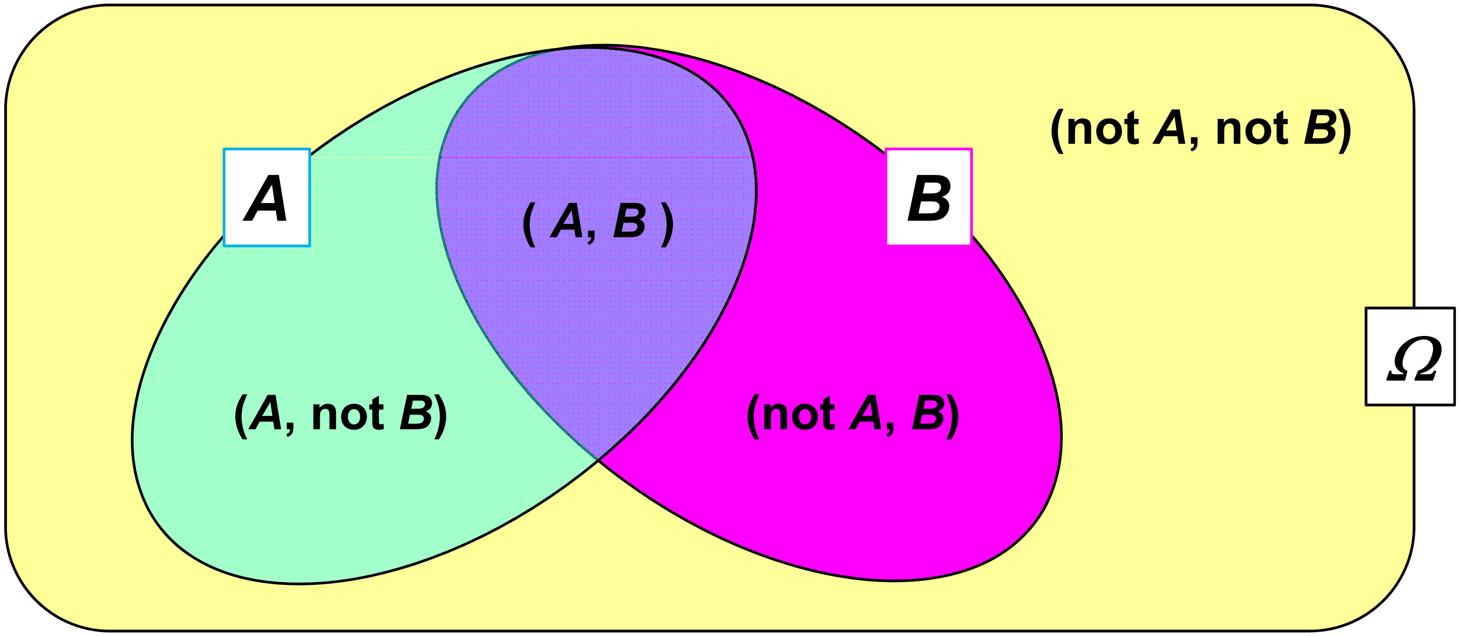
- Independence:
  - Outcomes on multiple rolls of a die
  - Outcomes on multiple flips of a coin
  - Height of two unrelated individuals
  - Probability of getting a king on successive draws from a deck, if card from each draw is *replaced*
- Dependence:
  - Height of two related individuals
  - Duration of successive eruptions of Old Faithful
  - Probability of getting a king on successive draws from a deck, if card from each draw is *not replaced*

Bayes rule

A way to find conditional probabilities for one variable when conditional probabilities for another variable are known.

$$p(B \mid A) = p(A \mid B) \cdot p(B) \,/\, p(A)$$

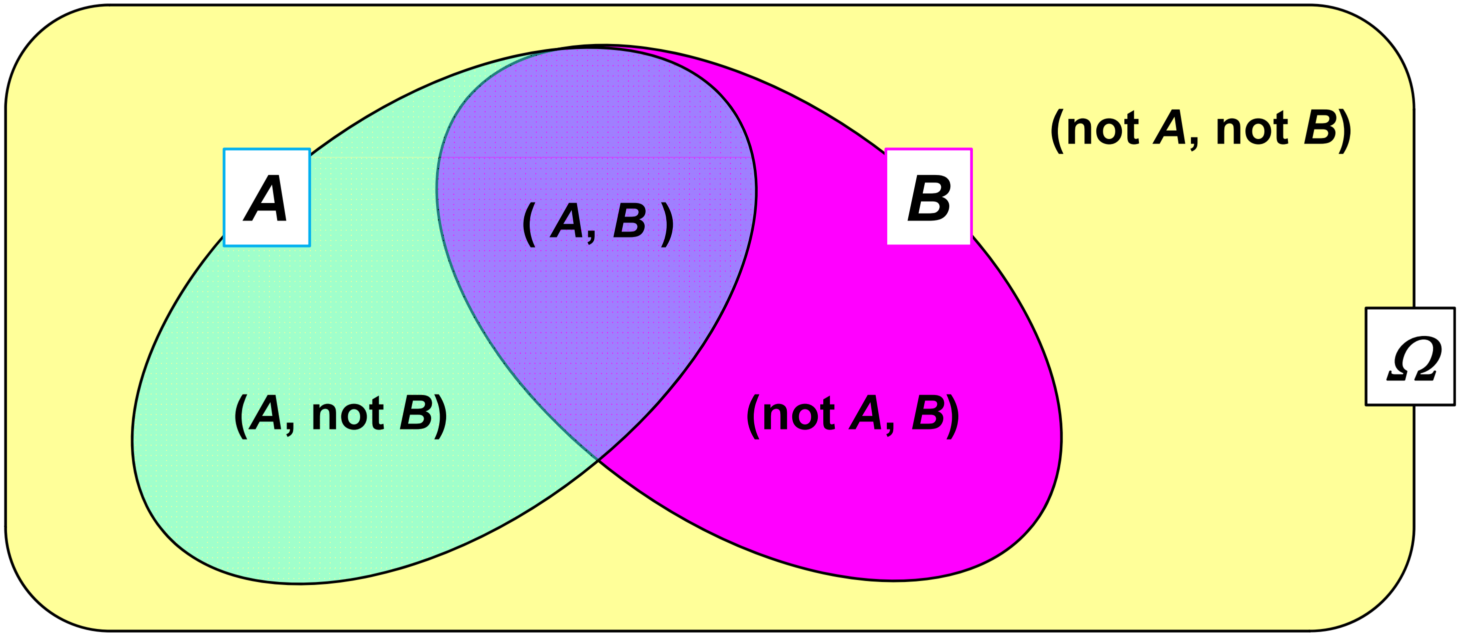
where $p(A) = p(A, B) + p(A, \text{not } B)$



Bayes rule

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

$$p(\,B \mid A\,) = p(\,A \mid B\,) \cdot p(\,B\,) \,/\, p(\,A\,)$$



Example of Bayes rule

- Marie is getting married tomorrow at an outdoor ceremony in the desert. In recent years, it has rained only 5 days each year. Unfortunately, the weatherman is forecasting rain for tomorrow. When it actually rains, the weatherman has forecast rain 90% of the time. When it doesn't rain, he has forecast rain 10% of the time. What is the probability it will rain on the day of Marie's wedding?
- Event *A*: The weatherman has forecast rain.
- Event *B*: It rains.
- We know:
  - $p(B) = 5 / 365 = 0.0137$ [It rains 5 days out of the year.]
  - $p(\text{not } B) = 360 / 365 = 0.9863$
  - $p(A \mid B) = 0.9$ [When it rains, the weatherman has forecast rain 90% of the time.]
  - $p(A \mid \text{not } B) = 0.1$ [When it does not rain, the weatherman has forecast rain 10% of the time.]

Example of Bayes rule, cont'd.

- We want to know $p(B \mid A)$, the probability it will rain on the day of Marie's wedding, given a forecast for rain by the weatherman. The answer can be determined from Bayes rule:

1. $p(B \mid A) = p(A \mid B) \cdot p(B) / p(A)$
2. $p(A) = p(A \mid B) \cdot p(B) + p(A \mid \text{not } B) \cdot p(\text{not } B) =$
   $(0.9)(0.014) + (0.1)(0.986) = 0.111$
3. $p(B \mid A) = (0.9)(0.0137) / 0.111 = 0.111$

- The result seems unintuitive but is correct. Even when the weatherman predicts rain, it only rains only about 11% of the time. Despite the weatherman's gloomy prediction, it is unlikely Marie will get rained on at her wedding.

Probabilities: when to add, when to multiply

- **ADD**: When you want to allow for occurrence of any of several possible outcomes of a *single* process. Comparable to logical OR.

- **MULTIPLY**: When you want to allow for simultaneous occurrence of *particular* outcomes from *more than one* process. Comparable to logical AND.
  - But only if the processes are *independent.*