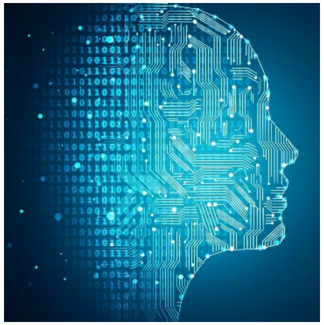
Principles of Machine Learning

Probability Review

Kevin Moon (kevin.moon@usu.edu)

STAT/CS 5810/6655



Motivation

- Machine learning methods require data to “learn” different tasks
- Data are typically collected in some random fashion
- Ideally, we want our machine learning algorithms to “generalize” to new, previously unseen data
- Given the randomness of the data, our ability to generalize includes some uncertainty
- Probability theory is useful for analyzing how this uncertainty affects the performance of machine learning methods
    - E.g., we can talk about a classifier’s probability of error

Possible sources of uncertainty

- Inherent stochasticity of the system being modeled
  - E.g. the quantum mechanics of electrons in a machine that takes measurements (i.e. noise)
- Incomplete observability
  - E.g. the Monty Hall problem: a game show contestant has to choose between three doors to win a prize. Two doors lead to a goat while a third leads to a car. From the contestant’s point of view, the outcome is uncertain.
- Incomplete modeling
  - E.g. discretization of continuous space

Outline

- Random variables (discrete and continuous)
- Properties of random variables
- Expectation and Variance
- Jointly distributed random variables
- Conditional distributions and independence
- Bayes Rule
- Covariance and Correlation
- Estimation theory

Random Variables

- (Informal definition) A **random variable** X is a variable whose value us unknown until some random experiment (i.e., measurement or observation) is conducted
- A possible value of X is called an **outcome**
- The set of all possible outcomes is called the **sample space**, often denoted Ω
- An **event** is a subset $A \subseteq \Omega$ to which a probability may be assigned
- Every random variable is associated to a **probability distribution** P which assigns probabilities to events.
- Two main types of random variables: **discrete** and **continuous**

Discrete Random Variables

- A random variable is *discrete* if its sample space is a discrete set

$$\Omega = \{x_1, x_2, \ldots\}$$

- The distribution of a discrete RV is defined by its *probability mass function* (pmf).
- A function $p : \Omega \to [0,1]$ is a valid pmf iff
  - $p(x_i) \geq 0$ for all i
  - $\sum_{x_i \in \Omega} p(x_i) = 1$
- The probability of an event A is given by

$$P(A) = \sum_{x_i \in A} p(x_i)$$

Examples of Discrete RVs

- **Example**: X = roll a fair, six-sided die
  - Random experiment = roll the die
  - $\Omega = \{1,2,3,4,5,6\}$
  - $p(k) = \frac{1}{6}, k \in \Omega$
- **Example**: X = roll a loaded, six-sided die where a 6 is twice as likely as the other outcomes
  - Random experiment = roll the die
  - $\Omega = \{1,2,3,4,5,6\}$
  - $p(k) = \begin{cases} \frac{1}{7} & \text{if } k \neq 6 \\ \frac{2}{7} & \text{if } k = 6 \end{cases}$

Discrete Uniform Distribution

- A discrete random variable has a **(discrete) uniform** distribution if its sample spaces is a finite set

$$\Omega = \{x_1, x_2, \dots, x_N\}$$

and its pmf is

$$p(x_i) = \frac{1}{N}$$

- **Examples**: Roll a fair die, flip a fair coin, etc.

Bernoulli Trials

- A random variable with sample space $\{0,1\}$ is called a **Bernoulli trial**.
- It's pmf is characterized by $p := P(\{1\})$, or the probability that $X = 1$.
- p is called the **success** probability
- **Example:** Roll a fair die and let $X = 1$ if a 5 turns up, and $X = 0$ otherwise.
  - Then $p = \frac{1}{6}$

Binomial Distribution

- We say X has a **binomial distribution** with parameters N and p if it is the sum of N independent Bernoulli trials with success probability p.
- Sample space:

$$\Omega = \{0,1,\dots,N\}$$

- pmf:

$$p(k) \coloneqq \binom{N}{k} p^k (1-p)^{N-k}, k \in \Omega$$

- **Example**: Roll a fair die 10 times, let X be the number of 5's observed. Then

$$X \sim binom\left(10, \frac{1}{6}\right)$$

Continuous Random Variables

- A random variable is **continuous** if its sample space is a continuum of points, i.e., an interval or union of intervals in $\mathbb{R}$
- The distribution of a continuous RV is defined by its **probability density function** (pdf)
- A function $p: \Omega \to \mathbb{R}$ is a valid pdf iff
  - $p(x) \geq 0$ for all $x \in \Omega$
  - $\int_\Omega p(x)dx = 1$
- The probability of an event A is given by

$$P(A) = \int_A p(x)dx$$

Continuous Uniform Distribution

- A continuous random variable has a **(continuous) uniform** distribution if its pdf is constant on the sample space.
- $\Omega = [a, b]$
- $p(x) = \begin{cases} \frac{1}{b-a}, & x \in \Omega \\ 0, & x \notin \Omega \end{cases}$
- **Example**: Draw an arrow on a Frisbee. Throw the Frisbee a long ways and let X be the bearing (wrt magnetic north) of the arrow after it lands.
  - $\Omega = [0,2\pi)$

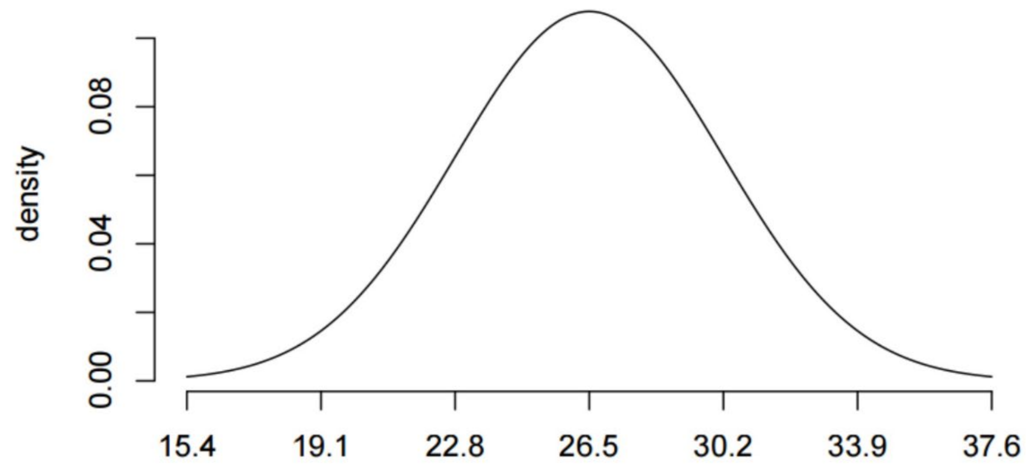
Gaussian Distribution

- A continuous random variable has a *Gaussian* or *normal* distribution if its pdf has the form

$$p(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

for some $\mu \in \mathbb{R}$ and $\sigma > 0$.

- **Examples:** Electronic noise, height of random person



Properties of Probability Distributions

- $P(\Omega) = 1$
- $P(A) \geq 0$ for all events A
- If $A_1, A_2, \ldots$ are disjoint, then

$$P(A_1 \cup A_2 \cup \cdots) = \sum_i P(A_i).$$

- $P(A^c) = 1 - P(A)$
- $P(\emptyset) = 0.$
- If $A \subseteq B$ then $P(A) \leq P(B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $\ldots$

Expectation

- The *expected value* of a random variable X is

$$E[X] := \sum_{x \in \Omega} x p(x)$$

if X is discrete, and

$$E[X] := \int_{\Omega} x p(x) \, dx$$

if X is continuous.

- Gives the average or mean of the probability distribution
- **Examples:**
  - $X \sim binom(N, p) \Rightarrow E[X] = Np$
  - $X \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow E[X] = \mu$

Variance

- The **variance** of a random variable X is

$$
\begin{aligned}
Var[X] &= E[(X - E[X])^2] \\
&= E[X^2] - (E[X])^2
\end{aligned}
$$

- Gives a measure of how much the probability distribution is spread about the mean

Jointly Distributed Random Variables

- Random variables X and Y are *jointly discrete* if they are both discrete and based on the same underlying random experiment.

- The sample space Ω is now the set of possible outcomes of the ordered pair (X, Y)

- The *joint pmf* of X and Y is the function

$$p(x, y) := \Pr(X = x \text{ and } Y = y)$$

For any event $A \subseteq \Omega$, the probability that $(X, Y) \in A$ is

$$\sum_{(x,y)\in A} p(x, y)$$

Group Exercise

Roll two fair six-sided dice and let $X = \max$, $Y = \min$.

1. What is the sample space?
2. Determine the joint pmf.
3. What is the probability that $Y \geq 4$?

Jointly Distributed Random Variables

- Informally, random variables X and Y are *jointly continuous* if they are both continuous and based on the same underlying random experiment.

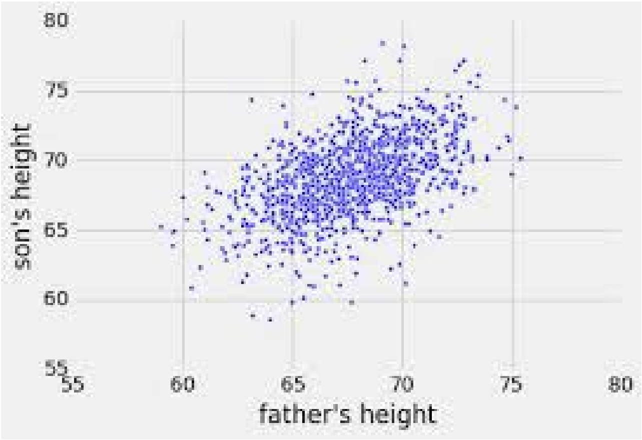
- Formally, X and Y are *jointly continuous* if there exists a function $p(x, y)$ (the joint pdf) such that, for all A, the probability of

$$(X, Y) \in A$$

is given by

$$\int_A p(x, y)\, dx dy$$

- **Example:** Bivariate Gaussian



Jointly Distributed Random Variables

- Natural generalization to multiple random variables

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_N \end{bmatrix}$$

- If $X_1, \ldots, X_N$ are jointly distributed, then each X_i is a (scalar) random variable whose pmf/pdf can be recovered from the joint pmf/pdf. For example, if X and Y are jointly continuous, then

$$p_X(x) = \int_{-\infty}^{\infty} p_{X,Y}(x,y)\, dy.$$

-

$$E[\boldsymbol{X}] := \begin{bmatrix} E[X_1] \\ \vdots \\ E[X_N] \end{bmatrix}$$

Conditional Distributions

- Suppose X and Y are jointly discrete, and $Y = y$ is observed. Then the *conditional distribution* of X given $Y = y$ is given by the conditional pmf
$$p_{X|Y}(x|y) := \frac{p_{XY}(x,y)}{p_Y(y)}.$$
- Suppose X and Y are jointly continuous, and $Y = y$ is observed. Then the *conditional distribution* of X given $Y = y$ is given by the conditional pdf
$$p_{X|Y}(x|y) := \frac{p_{XY}(x,y)}{p_Y(y)}.$$
- Natural extensions to multiple random variables, e.g., if $X_1, \ldots, X_5$ are jointly distributed, and X_4 and X_5 have already been observed, then
$$p(x_1, x_2, x_3|x_4, x_5) := \frac{p(x_1, \ldots, x_5)}{p(x_4, x_5)}$$

Independent Random Variables

- Jointly distributed random variables $X_1, \ldots, X_N$ are said to be *independent* if

$$p(x_1, x_2, \ldots, x_N) = p(x_1)p(x_2)\cdots p(x_N),$$

  i.e., "the joint is the product of marginals."

- Intuitively, if you know the outcome of some subset of independent RVs, it doesn't tell you anything about the other RVs.

- **Example:** Two consecutive flips of a coin

- For independent random variables, conditional distributions reduce to marginal distributions, e.g., if X and Y are independent, then

$$p(x|y) = \frac{p(x,y)}{p(y)} = \frac{p(x)p(y)}{p(y)} = p(x).$$

Law of Total Probability

- *Discrete case*: Let X, Y be jointly discrete. Then

$$\begin{aligned} p(x) &= \sum_y p(x, y) \\ &= \sum_y p(x|y)p(y) \end{aligned}$$

- *Continuous case*: Let X, Y be jointly continuous. Then

$$\begin{aligned} p(x) &= \int_{-\infty}^{\infty} p(x, y)\, dy \\ &= \int_{-\infty}^{\infty} p(x|y)p(y)\, dy \end{aligned}$$

Law of Total Expectation

- Denote by

$$E_{X|Y}[X|Y]$$

the expected value of X given Y, where Y is viewed as random.

- Thus $E_{X|Y}[X|Y]$ is random.
- LOTE:

$$E_X[X] = E_Y[E_{X|Y}[X|Y]].$$

Bayes Rule

- For jointly distributed X and Y,

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

Mixed Case

- What if X is continuous and Y is discrete? Can X and Y still be jointly distributed?
- Joint pmf/pdf no longer make sense.
- However, marginal and conditional distributions still make sense, so we can still calculate probabilities and expectations.
- **EXERCISE:** Suppose $Y \sim \text{Bernoulli}(1/3)$, $X|Y = 1 \sim \mathcal{N}(1, 1)$, and $X|Y = 0 \sim \mathcal{N}(0, 1)$. What is $\Pr(X \geq 0.5)$?

Covariance

- The **covariance** between two random variables X and Y is

$$\begin{aligned} Cov(X,Y) &= E[(X-E[X])(Y-E[Y])] \\ &= E[XY]-E[X]E[Y] \end{aligned}$$

- Measures how much the random variables vary together
  - May be negative
  - Zero if X and Y are independent
- Covariance matrices are also important:

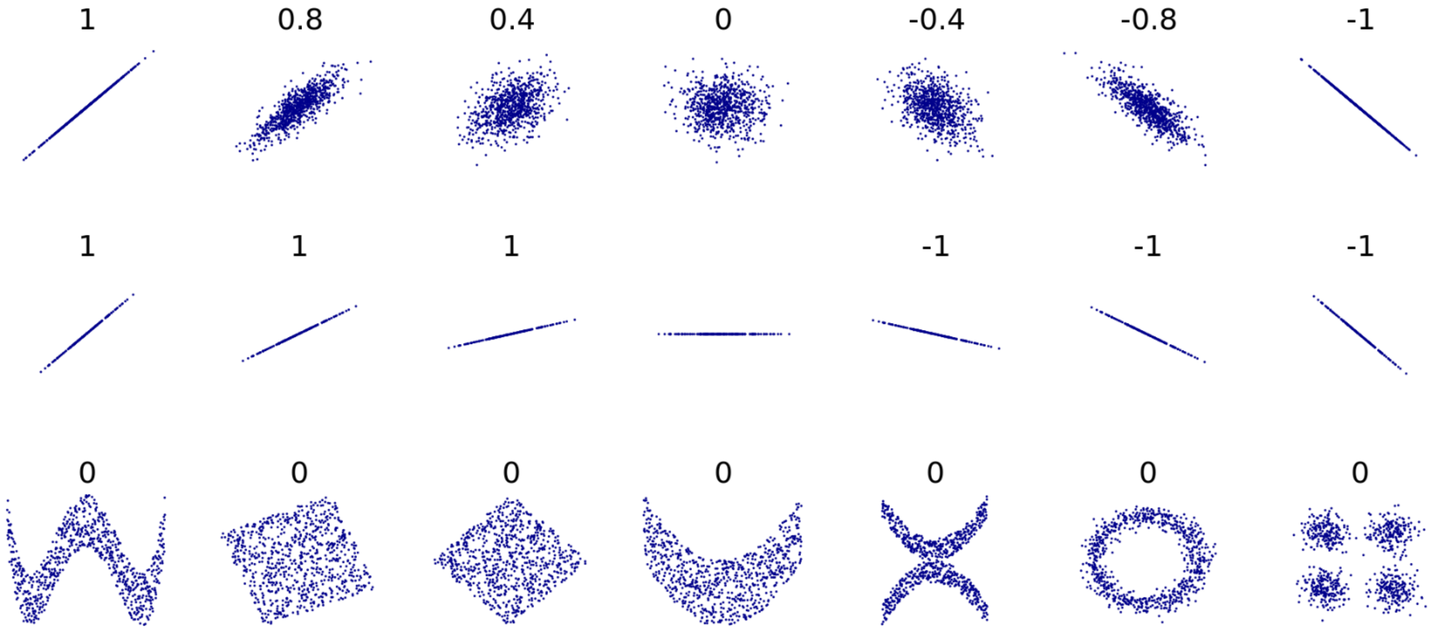
$$\Sigma_{\mathbf{X}} = \begin{bmatrix} Cov(X_1,X_1) & \dots & Cov(X_1,X_d) \\ \vdots & \ddots & \vdots \\ Cov(X_d,X_1) & \dots & Cov(X_d,X_d) \end{bmatrix}$$

Correlation

- **Pearson correlation coefficient**:

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var[X]Var[Y]}}$$

- Measures the strength of the linear relationship between X and Y



https://commons.wikimedia.org/w/index.php?curid=15165296

Properties of Expectation and Variance

Let $X_1, \ldots, X_n$ be a finite set of random variables and let $a_1, \ldots, a_n$ be scalars

- $E[\sum_{i=1}^{n} a_i X_i] = \sum_{i=1}^{n} a_i E[X_i]$
  - Not necessarily true in the limit
- $E[X_i X_j] = E[X_i] E[X_j]$ if X_i and X_j are independent
- $Var[a_i X_i] = a_i^2 Var[X_i]$
- $Var[\sum_{i=1}^{n} a_i X_i] = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j Cov(X_i, X_j)$
  - Note that $Cov(X_i, X_i) = Var[X_i]$
  - If all X_i and X_j are independent for $i \neq j$, the double sum reduces to $\sum_{i=1}^{n} a_i^2 Var[X_i]$
  - Also not necessarily true in the limit

An estimator

- Suppose we have data points $X_1, \dots, X_n$ drawn from a probability distribution (could be discrete or continuous) p
- Suppose p has some property or parameter c
  - Examples: mean, variance, success probability, etc.
- An estimator $\hat{c}(X_1, \dots, X_n)$ of c is a function of the data that attempts to estimate or approximate c
  - We usually just write $\hat{c}$ for short
- Example: the sample mean $\frac{1}{n}\sum_{i=1}^{n} X_i$ is an estimator of the mean of the probability distribution

Mean Squared Error

- How do we know if we have a good estimator?
- Measure its accuracy
- A common measure of accuracy is the mean squared error (MSE):

$$MSE(\hat{c}) = E[(\hat{c} - c)^2]$$

- The MSE can be decomposed into the sum of the variance of $\hat{c}$ and its squared **bias**
- Thus the MSE of an estimator can be minimized by minimizing the variance and the squared bias

Estimator Bias and Variance

- We already covered the formula for the variance: $E[(\hat{c} - E[\hat{c}])^2]$
- The bias of an estimator:

$$Bias[\hat{c}] = E[\hat{c}] - c$$

- Often, there is a tradeoff between bias and variance: decreasing the bias increases the variance and vice versa
  - A little more on this later in the course

Further Reading

- Review of Probability Theory on Canvas