CS 559 Machine Learning

Introduction and Overview

Yue Ning
Department of Computer Science
Stevens Institute of Technology

Instructor and TA Information

- Instructor: Yue Ning
  - Office hours: 2:00PM - 4:00PM Tuesday
  - Office:North Building 217
  - Email: Yue.Ning@stevens.edu
- Teaching Assistants:
  - Selcuk Karakas, fkarakas@stevens.edu
    Office hours: 11AM - 12PM Thursdays
  - Yujie Zhou, yzhou88@stevens.edu
    Office hours: 11AM - 12PM Mondays

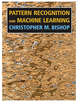
Course Information

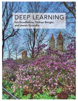
- Webpage: https://yue-ning.github.io/cs559-s19.html
- Prerequisite course
  - MA 222 Probability Theory
- Meeting
  - Time: 6:30 PM - 9:00 PM
  - Date: Wednesday
  - Location: BC 122
- Canvas: Announcements, Assignments, Discussions (Login to myStevens).

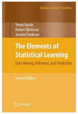
Reading

Not required, but recommended:

- Bishop, Christopher M., 2016. Pattern Recognition and Machine Learning. Springer-Verlag New York, Inc.
- Ian Goodfellow and Yoshua Bengio and Aaron Courville, 2016. Deep Learning, MIT Press.
- Hastie, Trevor and Tibshirani, Robert and Friedman, Jerome, 2001. The Elements of Statistical Learning. Springer New York Inc.







Course Prerequisites and Goals

Before this course, you should know......

- Programming (Python)
- Linear Algebra (Vector, Matrix, Projection, Eigenvalues/vector ...)
- Probability and Optimization (distributions, expectation/variance, objective function, ...)

Course Prerequisites and Goals

Before this course, you should know......

- Programming (Python)
- Linear Algebra (Vector, Matrix, Projection, Eigenvalues/vector ...)
- Probability and Optimization (distributions, expectation/variance, objective function, ...)

At the end of this course, you should......

- Be more proficient at math and programming
- Be able to recognize when and how a new problem can be solved with an existing technique
- Be able to implement general ML techniques for a variety of problem types

Alert

This class is not about:

- Introduction to machine learning tools/softwares

Coursework

Distribution

- Homework (40%)
- Midterm Exam (10%)
- Final Exam (15%)
- Course Project (25%)
- Participation (5%)
- Quizzes (5%)

Policy

- Late days: maximum of two(2) late days in total. If you use up your late days, late submission will not be graded.
- Canvas: ask questions on Canvas. Don't email directly.

Homework

- Goal: test your ability to understand knowledge of ML
- Mix of written and programming problems
- Submission: e-copy on Canvas
- Jupyter Notebook: it is recommended to use for your programming assignments.

Exams

- Goal: test your ability to use knowledge of ML to solve new problems
- All written problems. similar to written part of homework
- Closed book
- Covers all material up to the preceding work
- Mid-term and final: In the middle of the semester and the last week.
- Submission: In class and physical copy

Course Project

- Goal: select any problem you are interested and apply ML techniques to tackle it.
- Milestones: proposal report, final report, 3-min YouTube presentation!
- Task is open, please follow: task definition, implement baselines, evaluation, literature review, result analysis.
- Help: come to any office hours, TA, Internet.
- Submission: reports and code.

The Honor Code

- Collaborate and discuss together, but write code and reports independently.
- Do not look at classmates' writeup or code
- Do not share writeup/code (online and physical)
- Debugging: only look at input-output behavior
- Detect plagiarism.

Coursework grading

Distribution of (0-100)

- A $(90 - 100)$
- A- $(85 - 90)$
- B$+$ $(80 - 85)$
- B $(75 - 80)$
- B- $(70 - 75)$
- C$+$ $(65 - 70)$
- C $(60 - 65)$
- F (< 60)

Coursework grading

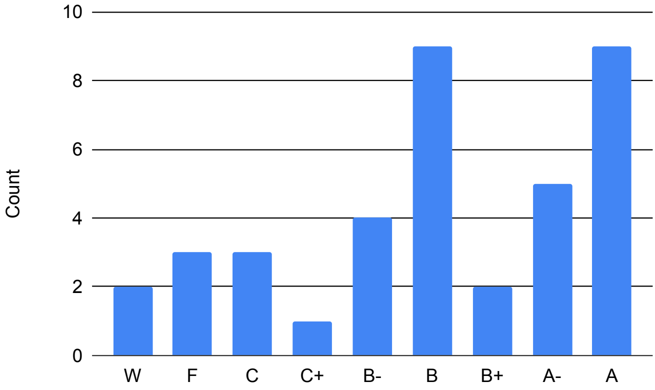


Figure: Students Performance Fall 2018

How to fail this course?

1. Do not submit the first homework assignment

2. Do not submit course project proposal/final report

3. Miss classes

4. Miss exams

AI and Machine Learning

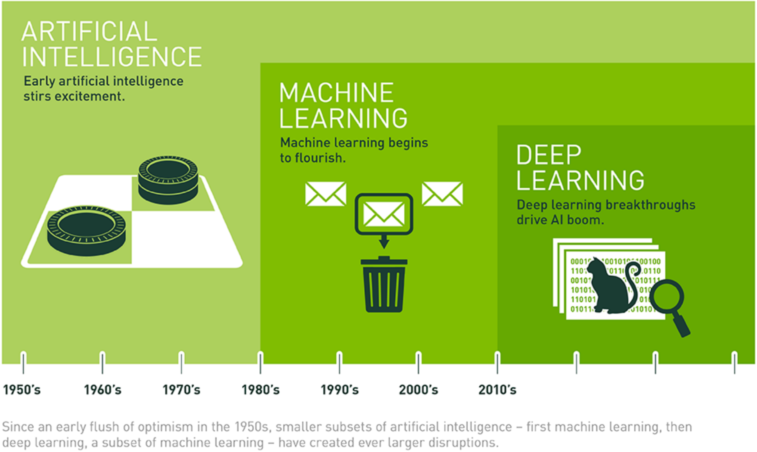
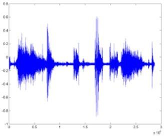


Figure: Nvidia blog

What is Machine Learning?

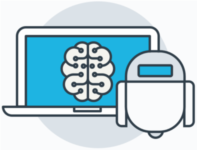
- Arthur Samuel (1959). Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.
- Automatically learn patterns from empirical data in order to improve their performance.
- An interdisciplinary (and relatively young) field focusing both on theoretical foundations of systems that learn, reason and act as well as on practical applications of these systems.

What is Machine Learning?



You said “Nice to meet you”

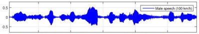
→



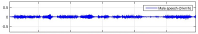
Program for learning

→

“How is it going?”



“Glad to see you”

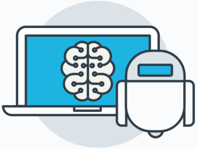


Testing audio data

What is Machine Learning?

This is “cat”

→



Program for learning

→

“Monkey”

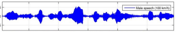
“Dog”

“Cat”

Testing image data

Machine Learning ≈ Looking for a function

- Speech recognition

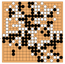
$f($ 

$) \rightarrow$ "Nice to meet you"

- Image recognition

$f($  $) \rightarrow$ "Cat"

- Playing Go

$f($  $) \rightarrow$ "5-5" (next move)

- Dialogue System

$f($"Hi"$) \rightarrow$ "Hello"

Training Framework

A set of functions $\rightarrow$ Goodness of function f $\leftarrow$ Training Data

Model $f_1, f_2, \ldots$

Pick the "best" function f^*

"Dog"

"Cat"

"Monkey

Testing Framework

 $\rightarrow$ **Using f^*** $\rightarrow$ “Cat”

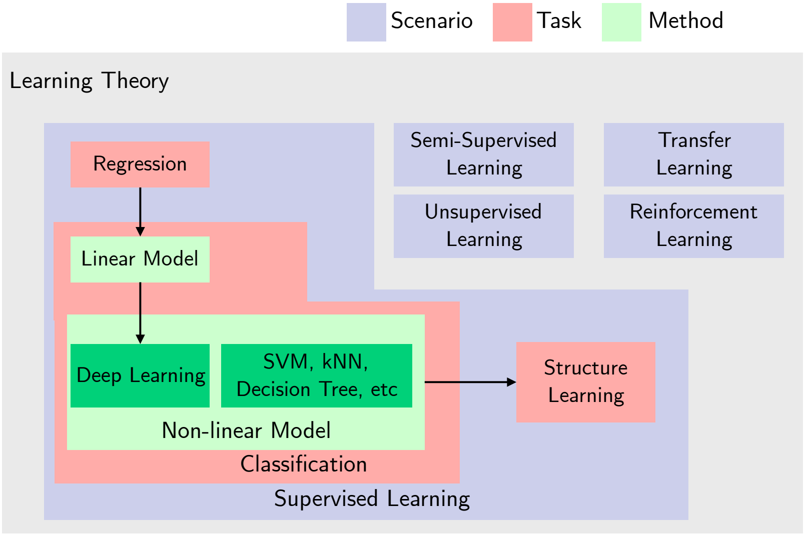
Many scientific fields

- Statistics: Inference from data, probabilistic models, learning theory,......
- Mathematics: Optimization theory, numerical methods, tools for theory,......
- Engineering: Signal processing, system identification, robotics, control, information theory, data-mining,.
- Economics: decision theory, operations research, econometrics,......
- Computer science: Artificial intelligence, computer vision, information retrieval, data-structures, implementations,......

Many scientific fields

- Physics: Energy minimization principles, entropy, capacity
- Psychology/Cognitive science: Computational linguistics, reinforcement learning, movement control,......
- Computational Neuroscience: Neural networks, principles of neural information
- Frequently: information flowing back in from applications domains, e.g. tools for bioinformatics getting used in other domains,......

Learning Map



Supervised Learning

- Sample data comprises input vectors along with the corresponding target values(labeled data).
- Supervised learning uses the given labeled data to find a model (hypothesis) that predicts the target values for previously unseen data.

Supervised Learning: Classification

Classification: each element in the sample is labeled as belonging to some class. There is no order among classes.

- Binary classification. (Examples: spam classification)

Input $\rightarrow$ $\boxed{\text{Function } f}$ $\rightarrow$ Yes (1) or No (0)

- Multi-calss classification. (Examples: document topic classification, image object classification, etc)

Input $\rightarrow$ $\boxed{\text{Function } f}$ $\rightarrow$ “Cat ” or “Dog” or “Monkey”

Classification - Deep Learning

- Image Recognition

 $\rightarrow X \rightarrow$

Function f

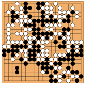
Hierarchical Structure

$\rightarrow$

Cat
Dog
......
Monkey

Classification - Deep Learning

- Playing Go



$\rightarrow X \rightarrow$ **Function f** $\rightarrow$

Hierarchical Structure

Next move
Each position is a class
(19*19 classes)

Supervised Learning: Regression

Regression: each element in the sample is associated with one or more continuous variables. Unlike classes, values have an order among them.

Example: predict house price

Location: Hoboken
Size: 1000sqft
School: [5,6,7]
......

$\rightarrow$ f $\rightarrow$ \$1.1M

Semi-supervised Learning

Unlabeled data may be easily available, while labeled data maybe expensive to obtain.
Semi-supervised learning: it exploits unlabeled examples, in addition to labeled ones, to improve the generalization ability of the resulting classifier.

- For example, recognizing cats and dogs

Transfer Learning

- For example, recognizing cats and dogs

Labelled data

"Cat"

"Dog"

Semi-Labelled data

"Monkey"

"Haruhi"

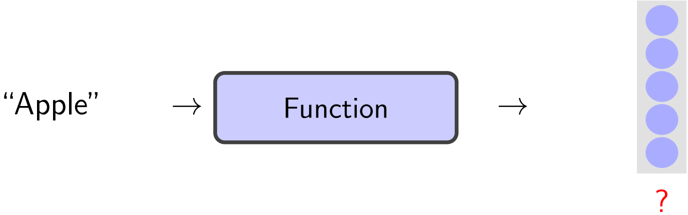
Data not related to the task considered (can be either labeled or unlabeled)

Unsupervised Learning

- The given data consists of input vectors without any corresponding target values.
- The goal is to discover groups of similar examples within the data (clustering), or to determine the distribution of data within the input space (density estimation).

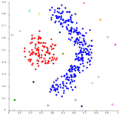
Unsupervised Learning

- Machine reading/understanding: learns the meaning of words from reading a lot of documents

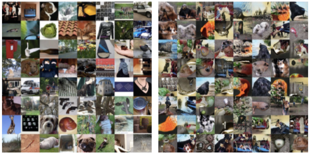


Unsupervised Learning

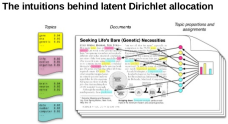
- Clustering



- Machine drawing [Salimans et al 2016]



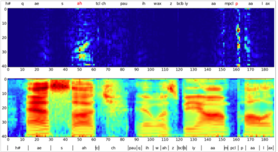
- Machine reading [Blei et al 2003]



[D. Blei, Probabilistic topic models, Communications of the ACM, 2012]

Structure Learning - Beyond Classification

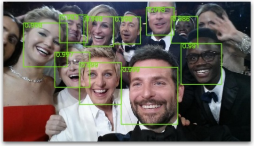
- Speech recognition [Graves et al 2013]



- Machine Translation [Google Brain 2016]

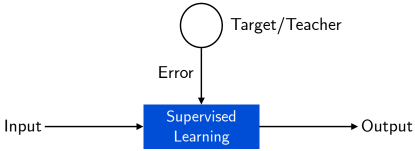
Input sentence:	***Translation (PBMT):***	***Translation (GNMT):***	***Translation (human):***
李克強此行將啟動中加總理年度對話機制，與加拿大總理杜魯多舉行兩國總理首次年度對話。	Li Keqiang premier added this line to start the annual dialogue mechanism with the Canadian Prime Minister Trudeau two prime ministers held its first annual session.	Li Keqiang will start the annual dialogue mechanism with Prime Minister Trudeau of Canada and hold the first annual dialogue between the two premiers.	Li Keqiang will initiate the annual dialogue mechanism between premiers of China and Canada during this visit, and hold the first annual dialogue with Premier Trudeau of Canada.

- Face recognition [Farfade et al 2015]

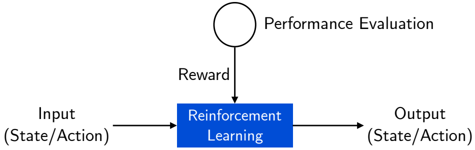


Reinforcement Learning - Beyond Classification

- Supervised: learning from teacher



- Reinforcement: learning from critics



Reinforcement Learning - Beyond Classification

- The problem here is to find suitable actions to take in a given situation in order to maximize a reward.
- Trial and error: no examples of optimal outputs are given.
- Trade-off between exploration (try new actions to see how effective they are) and exploitation (use actions that are known to give high reward).

Reinforcement Learning - Example

- Reinforcement

  First move $\rightarrow$ ......many moves......$\rightarrow$ Win!

Alpha Go is supervised learning + reinforcement learning

Quiz

Probability

Discrete Random Variable

- Sample space Ω: Possible "states" x of the random variable X
  (outcomes of the experiment, output of the system, measurement).

Discrete Random Variable

- Sample space Ω: Possible "states" x of the random variable X
  (outcomes of the experiment, output of the system, measurement).
- Discrete random variables either have a finite or countable number of states.

Discrete Random Variable

- Sample space Ω: Possible "states" x of the random variable X
  (outcomes of the experiment, output of the system, measurement).
- Discrete random variables either have a finite or countable number of states.
- Events: Possible combinations of states ('subsets of Ω')

Discrete Random Variable

- Probability mass function $P(X = x)$: A function which tells us how likely each possible outcome is.

$$P(X = x) = P_X(x) = P(x) \tag{1}$$

$$P(x) \geq 0 \text{ for each } x \tag{2}$$

$$\sum_{x \in \Omega} P(x) = 1 \tag{3}$$

$$P(A) = P(x \in A) = \sum_{x \in A} P(X = x) \tag{4}$$

- We write: $X|q \sim \text{Binomial}(q)$
- Bernoulli, Binomial, Multinonomial, Poisson

Conditional Probability

- Conditional probability: Recalculated probability of event A after someone tells you that event B happened.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{5}$$

$$P(A \cap B) = P(A|B)P(B) \tag{6}$$

Conditional Probability

- Conditional probability: Recalculated probability of event A after someone tells you that event B happened.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{5}$$

$$P(A \cap B) = P(A|B)P(B) \tag{6}$$

- Bayes Rule:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{7}$$

Expectation and Variance

- Expectation (or mean): $E(X) = \sum_x P(X = x)x$

Expectation and Variance

- Expectation (or mean): $E(X) = \sum_x P(X = x)x$
- Expectation of a function: $E(f(X)) = \sum_x P(X = x)f(x)$

Expectation and Variance

- Expectation (or mean): $E(X) = \sum_x P(X = x)x$
- Expectation of a function: $E(f(X)) = \sum_x P(X = x)f(x)$
- Moments = expectation of power of X: $M_k = E(X^k)$

Expectation and Variance

- Expectation (or mean): $E(X) = \sum_x P(X = x)x$
- Expectation of a function: $E(f(X)) = \sum_x P(X = x)f(x)$
- Moments = expectation of power of X: $M_k = E(X^k)$
- Variance: Average (squared) fluctuation from the mean

$$\begin{aligned} \mathrm{Var}(X) &= E((X - E(X))^2) && (8) \\ &= E(X^2) - E(X)^2 && (9) \\ &= M_2 - M_1^2 && (10) \end{aligned}$$

Expectation and Variance

- Expectation (or mean): $E(X) = \sum_x P(X = x)x$
- Expectation of a function: $E(f(X)) = \sum_x P(X = x)f(x)$
- Moments = expectation of power of X: $M_k = E(X^k)$
- Variance: Average (squared) fluctuation from the mean

$$\begin{aligned} \mathrm{Var}(X) &= E((X - E(X))^2) && (8) \\ &= E(X^2) - E(X)^2 && (9) \\ &= M_2 - M_1^2 && (10) \end{aligned}$$

- Standard deviation: Square root of variance.

Expectation and Variance

- Expectation (or mean): $E(X) = \sum_x P(X = x)x$
- Expectation of a function: $E(f(X)) = \sum_x P(X = x)f(x)$
- Moments = expectation of power of X: $M_k = E(X^k)$
- Variance: Average (squared) fluctuation from the mean

$$\begin{aligned} \mathrm{Var}(X) &= E((X - E(X))^2) && (8) \\ &= E(X^2) - E(X)^2 && (9) \\ &= M_2 - M_1^2 && (10) \end{aligned}$$

- Standard deviation: Square root of variance. Aside: Difference between expectation/variance of random variable and empirical average/variance.

Bivariate Distributions

- Joint distribution: $P(X = x, Y = y)$, a list of all probabilities of all possible pairs of observations

Bivariate Distributions

- Joint distribution: $P(X = x, Y = y)$, a list of all probabilities of all possible pairs of observations
- Marginal distribution: $P(X = x) = \sum_y P(X = x, Y = y)$

Bivariate Distributions

- Joint distribution: $P(X = x, Y = y)$, a list of all probabilities of all possible pairs of observations
- Marginal distribution: $P(X = x) = \sum_y P(X = x, Y = y)$
- Conditional distribution: $P(X = x|Y = y) = \frac{P(X=x,Y=y)}{P(y=y)}$

Bivariate Distributions

- Joint distribution: $P(X = x, Y = y)$, a list of all probabilities of all possible pairs of observations
- Marginal distribution: $P(X = x) = \sum_y P(X = x, Y = y)$
- Conditional distribution: $P(X = x|Y = y) = \frac{P(X=x,Y=y)}{P(y=y)}$
- $X|Y$ has distribution $P(X|Y)$, where $P(X|Y)$ specifies a "lookup-table" of all possible $P(X = x|Y = y)$

Bivariate Distributions

- Joint distribution: $P(X = x, Y = y)$, a list of all probabilities of all possible pairs of observations
- Marginal distribution: $P(X = x) = \sum_y P(X = x, Y = y)$
- Conditional distribution: $P(X = x | Y = y) = \frac{P(X=x, Y=y)}{P(y=y)}$
- $X|Y$ has distribution $P(X|Y)$, where $P(X|Y)$ specifies a "lookup-table" of all possible $P(X = x|Y = y)$

Conditioning and marginalization come up in Bayesian inference ALL the time: Condition on what you observe. Marginalize out the uncertainty.

Expectation and Covariance of Multivariate Distributions

- Conditional distributions are just distributions which have a (conditional) mean or variance.

Expectation and Covariance of Multivariate Distributions

- Conditional distributions are just distributions which have a (conditional) mean or variance.
- Covariance is the expected value of the product of fluctuations:

$$
\begin{aligned}
\mathrm{Cov}(X, Y) &= E\left((X - E(X))(Y - E(Y))\right) && (11)\\
&= E(XY) - E(X)E(Y) && (12)\\
\mathrm{Var}(\mathrm{X}) &= \mathrm{Cov}(X, X) && (13)
\end{aligned}
$$

Expectation and Covariance of Multivariate Distributions

- Conditional distributions are just distributions which have a (conditional) mean or variance.
- Covariance is the expected value of the product of fluctuations:

$$\begin{aligned} \mathrm{Cov}(X, Y) &= E\left((X - E(X))(Y - E(Y))\right) \qquad (11) \\ &= E(XY) - E(X)E(Y) \qquad (12) \\ \mathrm{Var(X)} &= \mathrm{Cov}(X, X) \qquad (13) \end{aligned}$$

Aside: One common way to construct bivariate random variables is to have a random variable whose parameter is another random variable.

Independence of Random Variables

- Intuitively, two events are independent if knowing that the first took places tells us nothing about the probability of the second: $P(A|B) = P(A)$

Independence of Random Variables

- Intuitively, two events are independent if knowing that the first took places tells us nothing about the probability of the second: $P(A|B) = P(A)$
- $P(A)P(B) = P(A \cap B)$

Independence of Random Variables

- Intuitively, two events are independent if knowing that the first took places tells us nothing about the probability of the second: $P(A|B) = P(A)$
- $P(A)P(B) = P(A \cap B)$
- Two random variables are independent if the joint p.m.f. is the product of the marginals: $P(X = x, Y = y) = P(X = x)P(Y = y)$.
- If X and Y are independent, we write $X \perp Y$. Knowing the value of X does not tell us anything about Y.

Independence of Random Variables

- Intuitively, two events are independent if knowing that the first took places tells us nothing about the probability of the second: $P(A|B) = P(A)$
- $P(A)P(B) = P(A \cap B)$
- Two random variables are independent if the joint p.m.f. is the product of the marginals: $P(X = x, Y = y) = P(X = x)P(Y = y)$.
- If X and Y are independent, we write $X \perp Y$. Knowing the value of X does not tell us anything about Y.
- If X and Y are independent, $\mathrm{Cov}(X, Y) = 0$.

Independence of Random Variables

- Intuitively, two events are independent if knowing that the first took places tells us nothing about the probability of the second: $P(A|B) = P(A)$
- $P(A)P(B) = P(A \cap B)$
- Two random variables are independent if the joint p.m.f. is the product of the marginals: $P(X = x, Y = y) = P(X = x)P(Y = y)$.
- If X and Y are independent, we write $X \perp Y$. Knowing the value of X does not tell us anything about Y.
- If X and Y are independent, $\mathrm{Cov}(X, Y) = 0$.

Aside: Mutual information is a measure of how "non-independent" two random variables are.

Multivariate Distributions

- $\mathbf{X}, \mathbf{x}$ are vector valued.
- Mean: $E(\mathbf{X}) = \sum_{\mathbf{x}} \mathbf{x} P(\mathbf{x})$
- Covariance matrix:

$$\mathrm{Cov}(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j) \tag{14}$$

$$\mathrm{Cov}(\mathbf{X}) = E(\mathbf{X}\mathbf{X}^\top) - E(\mathbf{X})E(\mathbf{X})^\top \tag{15}$$

Multivariate Distributions

- $\mathbf{X}, \mathbf{x}$ are vector valued.
- Mean: $E(\mathbf{X}) = \sum_{\mathbf{x}} \mathbf{x} P(\mathbf{x})$
- Covariance matrix:

$$\mathrm{Cov}(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j) \tag{14}$$

$$\mathrm{Cov}(\mathbf{X}) = E(\mathbf{X}\mathbf{X}^\top) - E(\mathbf{X})E(\mathbf{X})^\top \tag{15}$$

- Conditional and marginal distributions: Can define and calculate any (multi or single-dimensional) marginals or conditional distributions we need: $P(X_1)$, $P(X_1, X_2)$, $P(X_1, X_2, X_3 | X_4)$, etc..

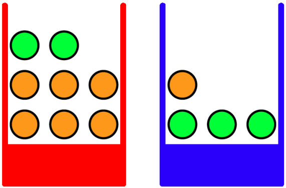
Example (from Bishop. [2006])

Assuming, we know that:

$$P(B = r) = 4/10 \quad (16)$$
$$P(B = b) = 6/10 \quad (17)$$

The probability of selecting a fruit from a given box is:

$$P(F = a|B = r) = 1/4 \quad (18)$$
$$P(F = o|B = r) = 3/4 \quad (19)$$
$$P(F = a|B = b) = 3/4 \quad (20)$$
$$P(F = o|B = b) = 1/4 \quad (21)$$



Example (from Bishop. [2006])

- What is the probability of choosing an apple?

Example (from Bishop. [2006])

- What is the probability of choosing an apple?
- Conditional probability:

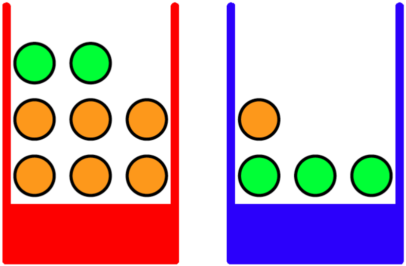
$$
\begin{aligned}
P(F = a) = P(F = a|B = r)P(B = r) \\
+ P(F = a|B = b)P(B = b) \qquad (22) \\
= \frac{1}{4} \times \frac{4}{10} + \frac{3}{4} \times \frac{6}{10} = 11/20 \qquad (23)
\end{aligned}
$$

Example (from Bishop. [2006])

- What is the probability of choosing an apple?
- Conditional probability:

$$
\begin{aligned}
P(F=a) &= P(F=a|B=r)P(B=r) \\
&\quad + P(F=a|B=b)P(B=b) && (22) \\
&= \frac{1}{4} \times \frac{4}{10} + \frac{3}{4} \times \frac{6}{10} = 11/20 && (23)
\end{aligned}
$$

- Thus, the probability of choosing an orange is $P(F=o) = 1 - 11/20 = 9/20$



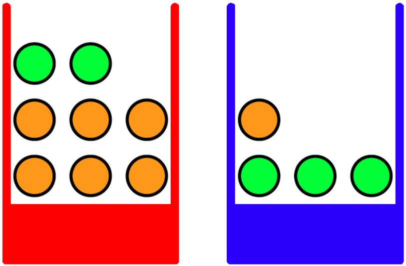
Example (from Bishop. [2006])

We are told that a piece of fruit has been selected and it is an orange, and we would like to know which box it came from.

Example (from Bishop. [2006])

We are told that a piece of fruit has been selected and it is an orange, and we would like to know which box it came from. Using Bayes' theorem,

$$P(B = r|F = o) = \frac{P(F = o|B = r)P(B = r)}{P(F = o)} = \frac{\frac{3}{4} \times \frac{4}{10}}{\frac{9}{20}} = 2/3 \tag{24}$$



Prior vs. Posterior

Prior Probability

If we had been asked which box had been chosen before being told the identity of the selected item of fruit, then the most complete information we have available is provided by the probability $P(B)$.

Posterior Probability

Once we are told that the fruit is an orange, we can then use Bayes' theorem to compute the probability $P(B|F)$, which we shall call the posterior probability because it is the probability obtained after we have observed F.

Bayesian Probabilities

Bayesian view: probabilities provide a quantification of uncertainty. Before observing the data, the assumptions about w are captured in the form of a prior probability distribution $P(\mathbf{w})$. The effect of the observed data $\mathcal{D} = \{(x_1, y_1), ..., (x_N, y_N)\}$ is expressed by $P(\mathcal{D}|\mathbf{w})$.

Bayes' theorem:

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

Bayes' theorem in words:

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

Continuous random variables

- A random variable X is continuous if its sample space X is uncountable.
- In this case, $P(X = x) = 0$ for each x.

Continuous random variables

- A random variable X is continuous if its sample space X is uncountable.
- In this case, $P(X = x) = 0$ for each x.
- If $p_X(x)$ is a probability density function for X, then

$$P(a < X < b) = \int_a^b p(x)dx \tag{25}$$

$$P(a < X < a + dx) \approx p(a) \cdot dx \tag{26}$$

Continuous random variables

- A random variable X is continuous if its sample space X is uncountable.
- In this case, $P(X = x) = 0$ for each x.
- If $p_X(x)$ is a probability density function for X, then

$$P(a < X < b) = \int_a^b p(x)dx \qquad (25)$$

$$P(a < X < a + dx) \approx p(a) \cdot dx \qquad (26)$$

- The cumulative distribution function is $F_X(x) = P(X < x)$. We have that $p_X(x) = F'(x)$, and $F(x) = \int_{-\infty}^{x} p(s)ds$.

Continuous random variables

- A random variable X is continuous if its sample space X is uncountable.
- In this case, $P(X = x) = 0$ for each x.
- If $p_X(x)$ is a probability density function for X, then

$$P(a < X < b) = \int_a^b p(x)dx \qquad (25)$$

$$P(a < X < a + dx) \approx p(a) \cdot dx \qquad (26)$$

- The cumulative distribution function is $F_X(x) = P(X < x)$. We have that $p_X(x) = F'(x)$, and $F(x) = \int_{-\infty}^{x} p(s)ds$.
- More generally: If A is an event, then

$$P(A) = P(X \in A) = \int_{x \in A} p(x)dx \qquad (27)$$

$$P(\Omega) = P(X \in \Omega) = \int_{x \in \Omega} p(x)dx = 1 \qquad (28)$$

Two sloppy facts: Probability vs. Probability Density

- $P(X = x)$: "the probability of X" when they really mean "the probability density of X evaluated at x".

Two sloppy facts: Probability vs. Probability Density

- $P(X = x)$: "the probability of X" when they really mean "the probability density of X evaluated at x".
- "We need to integrate out X": when X are discrete random variables, the integrals would need to be replaced by sums.

Two sloppy facts: Probability vs. Probability Density

- $P(X = x)$: "the probability of X" when they really mean "the probability density of X evaluated at x".
- "We need to integrate out X": when X are discrete random variables, the integrals would need to be replaced by sums.
- It is usually clear from the context whether a random variable is discrete or continuous.

Mean, Variance and Conditionals

- Mean: $E(X) = \int_x x \cdot p(x) dx$
- Variance: $\mathrm{Var}(X) = E(X^2) - E(X)^2$
- Example: Uniform [quiz]

Mean, Variance and Conditionals

- Mean: $E(X) = \int_x x \cdot p(x)dx$
- Variance: $\text{Var}(X) = E(X^2) - E(X)^2$
- Example: Uniform [quiz]
- If X has pdf $p(x)$, then $X|(X \in A)$ has pdf

$$p_{X|A}(x) = \frac{p(x)}{P(A)} = \frac{p(x)}{\int_{x\in A} p(x)dx} \tag{29}$$

Mean, Variance and Conditionals

- Mean: $E(X) = \int_x x \cdot p(x) dx$
- Variance: $\text{Var}(X) = E(X^2) - E(X)^2$
- Example: Uniform [quiz]
- If X has pdf $p(x)$, then $X|(X \in A)$ has pdf

$$p_{X|A}(x) = \frac{p(x)}{P(A)} = \frac{p(x)}{\int_{x \in A} p(x) dx} \tag{29}$$

- Only makes sense if $P(A) > 0$!

Bivariate Continuous Distributions

- $p_{X,Y}(x, y)$, joint probability density function of X and Y
- $\int_x \int_y p(x, y) dxdy = 1$

Bivariate Continuous Distributions

- $p_{X,Y}(x, y)$, joint probability density function of X and Y
- $\int_x \int_y p(x, y)dxdy = 1$
- Marginal distribution: $p(x) = \int_{-\infty}^{\infty} p(x, y)dy$

Bivariate Continuous Distributions

- $p_{X,Y}(x,y)$, joint probability density function of X and Y
- $\int_x \int_y p(x,y)dxdy = 1$
- Marginal distribution: $p(x) = \int_{-\infty}^{\infty} p(x,y)dy$
- Conditional distribution: $p(x|y) = \frac{p(x,y)}{p(y)}$
- Note: $P(Y = y) = 0!$
- Independence: X and Y are independent if $p_{X,Y}(x,y) = p_X(x)p_Y(y)$

The Univariate Gaussian

- Probability density function

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \tag{30}$$

- Easy to validate:

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma^2)dx = 1 \tag{31}$$

- Expectation

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma^2)xdx = \mu \tag{32}$$

- Variance

$$var[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2 \tag{33}$$

Products of Gaussian pdfs

- Suppose $p_1(x) = \mathcal{N}(x, \mu_1, 1/\beta_1)$ and $p_2(x) = \mathcal{N}(x, \mu_2, 1/\beta_2)$, then

Products of Gaussian pdfs

- Suppose $p_1(x) = \mathcal{N}(x, \mu_1, 1/\beta_1)$ and $p_2(x) = \mathcal{N}(x, \mu_2, 1/\beta_2)$, then

$$p_1(x)p_2(x) \propto \mathcal{N}(x, \mu, 1/\beta) \tag{34}$$

$$\beta = \beta_1 + \beta_2 \tag{35}$$

$$\mu = \frac{1}{\beta}(\beta_1\mu_1 + \beta_2\mu_2) \tag{36}$$

Products of Gaussian pdfs

- Suppose $p_1(x) = \mathcal{N}(x, \mu_1, 1/\beta_1)$ and $p_2(x) = \mathcal{N}(x, \mu_2, 1/\beta_2)$, then

$$p_1(x)p_2(x) \propto \mathcal{N}(x, \mu, 1/\beta) \tag{34}$$

$$\beta = \beta_1 + \beta_2 \tag{35}$$

$$\mu = \frac{1}{\beta}(\beta_1\mu_1 + \beta_2\mu_2) \tag{36}$$

In general:

$$p_1(x)p_2(x)...p_n(x) \propto \mathcal{N}(x, \mu, 1/\beta) \tag{37}$$

$$\beta = \sum_n \beta_n \tag{38}$$

$$\mu = \frac{1}{\beta}\sum_n \mu_n\beta_n \tag{39}$$

Products of Gaussian pdfs

- Suppose $p_1(x) = \mathcal{N}(x, \mu_1, 1/\beta_1)$ and $p_2(x) = \mathcal{N}(x, \mu_2, 1/\beta_2)$, then

$$p_1(x)p_2(x) \propto \mathcal{N}(x, \mu, 1/\beta) \tag{34}$$
$$\beta = \beta_1 + \beta_2 \tag{35}$$
$$\mu = \frac{1}{\beta}(\beta_1\mu_1 + \beta_2\mu_2) \tag{36}$$

In general:

$$p_1(x)p_2(x)...p_n(x) \propto \mathcal{N}(x, \mu, 1/\beta) \tag{37}$$
$$\beta = \sum_n \beta_n \tag{38}$$
$$\mu = \frac{1}{\beta}\sum_n \mu_n\beta_n \tag{39}$$

This is also true for multivariate Gaussians!

ML estimator

- Assuming data points are independent and identically distributed (i.i.d.), the probability of the data set given μ and σ^2 (the likelihood function):

$$P(\mathbf{x}|\mu,\sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu,\sigma^2) \tag{40}$$

- Log-likelihood:

$$\log P(\mathbf{x}|\mu,\sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\log\sigma^2 - \frac{N}{2}\log(2\pi) \tag{41}$$

- Maximizing Log-likelihood with respect to μ and σ^2:

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n, \quad \sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{ML})^2 \tag{42}$$

ML estimator

- The ML solutions μ_{ML} and σ^2_{ML} are functions of the data set values $x_1, .., x_N$. The expectations of these quantities w.r.t the data set values:

$$\mathbb{E}[\mu_{ML}] = \mu \tag{43}$$

$$\mathbb{E}[\sigma^2_{ML}] = \left(\frac{N-1}{N}\right)\sigma^2 \tag{44}$$

- The ML estimator obtains correct means but underestimate the true variance by a factor $\frac{N-1}{N}$

MAP estimator

- Given input values $\mathbf{x} = (x_1, ..., x_N)^T$ and their corresponding target values $\mathbf{y} = (y_1, ..., y_N)^T$.
- Express our uncertainty over the values of the target variables:

$$p(y|x, \mathbf{w}, \beta) = \mathcal{N}(y|f(x, \mathbf{w}), \beta^{-1})$$

- Using our training data to determine the unknown parameters $\mathbf{w}, \beta$ by maximum likelihood:

$$p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(y_n|f(x_n, \mathbf{w}), \beta^{-1})$$

- Log Likelihood:

$$lnp(y|x, \mathbf{w}, \beta) = -\frac{\beta}{2}\sum_{n=1}^{N}(f(x_n, \mathbf{w}) - y_n)^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi)$$

MAP estimator

- A more Bayesian approach by introducing a prior distribution for $\mathbf{w}$:

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(\mathbf{M}+1)/2} \exp\left(-\frac{\alpha}{2}\mathbf{w}^{\mathbf{T}}\mathbf{w}\right)$$

- Using Bayes' theorm, the posterior distribution for w:

$$p(\mathbf{w}|\mathbf{x}, \mathbf{y}, \alpha, \beta) \propto p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\alpha)$$

- Taking the negative logarithm, Maximum posterior (MAP) is equivalent to minimizing the regularized sum-of-squares error function:

$$\frac{\beta}{2} \sum_{n=1}^{N} \{f(x_n, \mathbf{w}) - y_n)\}^2 + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} \tag{45}$$

Bayesian Probabilities

- A key issue in pattern recognition is uncertainty. It is due to incomplete and/or ambiguous information, i.e. finite and noisy data.
- Probability theory and decision theory provide the tools to make optimal predictions given the limited available information.
- In particular, the Bayesian interpretation of probability allows to quantify uncertainty, and make precise revisions of uncertainty in light of new evidence.