Probability, Decision Theory, and Loss Functions

CMSC 678
UMBC

Some slides adapted from Hamed Pirsiavash
Logistics Recap

Piazza (ask & answer questions):
https://piazza.com/umbc/spring2019/cmsc678

Course site:
https://www.csee.umbc.edu/courses/graduate/678/spring19

Evaluation submission site:
https://www.csee.umbc.edu/courses/graduate/678/spring19/submit
Course Announcement: Assignment 1

Due Friday, 2/8 (~9 days)

Math & programming review

Discuss with others, but write, implement and complete on your own
A Terminology Buffet

Classification

Regression

Clustering

the task: what kind of problem are you solving?
A Terminology Buffet

**Classification**
- Regression

**Clustering**
- Fully-supervised
- Semi-supervised
- Un-supervised

*the task*: what kind of problem are you solving?

*the data*: amount of human input/number of labeled examples
A Terminology Buffet

Classification

Regression

Clustering

Fully-supervised

Semi-supervised

Un-supervised

the **task**: what kind of problem are you solving?

the **data**: amount of human input/number of labeled examples

the **approach**: how any data are being used

Probabilistic

Neural

Generative

Memory-based

Conditional

Exemplar

Spectral

...
Outline

Review+Extension

Probability

Decision Theory

Loss Functions
What does it mean to learn?

Generalization
Machine Learning Framework: Learning

instances are typically examined independently

scoring model $\text{score}_\theta(X)$

Machine Learning Predictor

evaluator

Gold/correct labels

$F(\theta)$ objective

give feedback to the predictor
Model, parameters and hyperparameters

Model: mathematical formulation of system (e.g., classifier)

Parameters: primary “knobs” of the model that are set by a learning algorithm

Hyperparameter: secondary “knobs”
Gradient Ascent

\[ \arg \max_{\theta} F(\theta) \]
General ML Consideration: Inductive Bias

What do we know before we see the data, and how does that influence our modeling decisions?
General ML Consideration: Inductive Bias

What do we know *before* we see the data, and how does that influence our modeling decisions?

Partition these into two groups...
General ML Consideration: Inductive Bias

What do we know \textit{before} we see the data, and how does that influence our modeling decisions?

Partition these into two groups

Who selected \textit{red} vs. \textit{blue}?
General ML Consideration: Inductive Bias

What do we know *before* we see the data, and how does that influence our modeling decisions?

Partition these into two groups

Who selected *red* vs. *blue*?

Who selected *●* vs. *▲*?
General ML Consideration: Inductive Bias

What do we know *before* we see the data, and how does that influence our modeling decisions?

Partition these into two groups

Who selected *red* vs. *blue*?

Who selected *black* vs. *triangle*?

Tip: Remember how your own biases/interpretation are influencing your approach

Courtesy Hamed Pirsiavash
Today’s Goals:

1. Remember Probability/Statistics

2. Understand Optimizing Empirical Risk
Outline

Review+Extension

Probability

Decision Theory

Loss Functions
Probability Prerequisites

- Basic probability axioms and definitions
- Joint probability
- Probabilistic Independence
- Marginal probability
- Definition of conditional probability
- Bayes rule
- Probability chain rule
- Common distributions
- Expected Value (of a function) of a Random Variable
(Most) Probability Axioms

\[ p(\text{everything}) = 1 \]
\[ p(\emptyset) = 0 \]
\[ p(A) \leq p(B), \text{ when } A \subseteq B \]
\[ p(A \cup B) = p(A) + p(B), \text{ when } A \cap B = \emptyset \]

\[ p(A \cup B) \neq p(A) + p(B) \]
\[ p(A \cup B) = p(A) + p(B) - p(A \cap B) \]
Probabilities and Random Variables

Random variables: variables that represent the possible outcomes of some random “process”
Probabilities and Random Variables

Random variables: variables that represent the possible outcomes of some random “process”

Example #1: A (weighted) coin that can come up heads or tails

X is a random variable denoting the possible outcomes
X=HEADS or X=TAILS
Probabilities and Random Variables

Random variables: variables that represent the possible outcomes of some random “process”

Example #1: A (weighted) coin that can come up heads or tails
  X is a random variable denoting the possible outcomes
  X=HEADS or X=TAILS

Example #2: Measuring the amount of snow that fell in the last storm
  Y is a random variable denoting the amount snow that fell, in inches
  Y=0, or Y=0.5, or Y=1.0495928591, or Y=10, or ...
Probabilities and Random Variables

Random variables: variables that represent the possible outcomes of some random “process”

Example #1: A (weighted) coin that can come up heads or tails
   X is a random variable denoting the possible outcomes
   X=HEADS or X=TAILS  
   DISCRETE random variable

Example #2: Measuring the amount of snow that fell in the last storm
   Y is a random variable denoting the amount snow that fell, in inches
   Y=0, or Y=0.5, or Y=1.0495928591, or Y=10, or ...
   CONTINUOUS random variable
## Random Variables

<table>
<thead>
<tr>
<th>If X is a...</th>
<th>Discrete random variable</th>
<th>Continuous random variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>The values ( k ) that ( X ) can take are</td>
<td>Discrete: finite or countably infinite (e.g., integers)</td>
<td>Continuous: uncountably infinite (e.g., real values)</td>
</tr>
</tbody>
</table>
## Random Variables

<table>
<thead>
<tr>
<th>If X is a...</th>
<th>Discrete random variable</th>
<th>Continuous random variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>The values ( k ) that ( X ) can take are</td>
<td>Discrete: finite or countably infinite (e.g., integers)</td>
<td>Continuous: uncountably infinite (e.g., real values)</td>
</tr>
<tr>
<td>The function that gives the relative likelihood of a value ( p(X=k) ) is a</td>
<td>probability mass function (PMF)</td>
<td>probability density function (PDF)</td>
</tr>
</tbody>
</table>
# Random Variables

<table>
<thead>
<tr>
<th></th>
<th>Discrete random variable</th>
<th>Continuous random variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>The values $k$ that $X$ can take are</td>
<td>Discrete: finite or countably infinite (e.g., integers)</td>
<td>Continuous: uncountably infinite (e.g., real values)</td>
</tr>
<tr>
<td>The function that gives the relative likelihood of a value $p(X=k)$ is a</td>
<td>probability mass function (PMF)</td>
<td>probability density function (PDF)</td>
</tr>
<tr>
<td>The values that PMF/PDF can take are</td>
<td>$0 \leq p(X=k) \leq 1$</td>
<td>$p(X=k) \geq 0$</td>
</tr>
</tbody>
</table>
## Random Variables

<table>
<thead>
<tr>
<th>If X is a...</th>
<th>Discrete random variable</th>
<th>Continuous random variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>The values k that X can take are</td>
<td>Discrete: finite or countably infinite (e.g., integers)</td>
<td>Continuous: uncountably infinite (e.g., real values)</td>
</tr>
<tr>
<td>The function that gives the relative likelihood of a value $p(X=k)$ is a</td>
<td>probability mass function (PMF)</td>
<td>probability density function (PDF)</td>
</tr>
<tr>
<td>The values that PMF/PDF can take are</td>
<td>$0 \leq p(X=k) \leq 1$</td>
<td>$p(X=k) \geq 0$</td>
</tr>
<tr>
<td>We “add” with</td>
<td>Sums ($\sum$)</td>
<td>Integrals ($\int$)</td>
</tr>
</tbody>
</table>
# Random Variables

If $X$ is a...

<table>
<thead>
<tr>
<th>The values $k$ that $X$ can take are</th>
<th>Discrete random variable</th>
<th>Continuous random variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discrete: finite or countably infinite (e.g., integers)</td>
<td>Continuous: uncountably infinite (e.g., real values)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>The function that gives the relative likelihood of a value $p(X=k)$ is a</th>
<th>probability mass function (PMF)</th>
<th>probability density function (PDF)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>The values that PMF/PDF can take are</th>
<th>$0 \leq p(X=k) \leq 1$</th>
<th>$p(X=k) \geq 0$</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>We “add” with</th>
<th>Sums ($\sum$)</th>
<th>Integrals ($\int$)</th>
</tr>
</thead>
</table>

| Our PMF/PDF satisfies $p($everything$)=1$ by | $\sum_{k} p(X = k) = 1$ | $\int p(x)dx = 1$ |
Probability Prerequisites

Basic probability axioms and definitions

Joint probability

Probabilistic Independence

Marginal probability

Definition of conditional probability

Bayes rule

Probability chain rule

Common distributions

Expected Value (of a function) of a Random Variable
Joint Probability

Probability that multiple things “happen together”
Joint Probability

Probability that multiple things “happen together”

\[ p(x,y), \ p(x,y,z), \ p(x,y,w,z) \]

Symmetric: \( p(x,y) = p(y,x) \)
Joint Probability

Probability that multiple things “happen together”

\[ p(x,y), p(x,y,z), p(x,y,w,z) \]

Symmetric: \( p(x,y) = p(y,x) \)

Form a table based of outcomes: sum across cells = 1

<table>
<thead>
<tr>
<th>( p(x,y) )</th>
<th>( Y=0 )</th>
<th>( Y=1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>X=“cat”</td>
<td>.04</td>
<td>.32</td>
</tr>
<tr>
<td>X=“dog”</td>
<td>.2</td>
<td>.04</td>
</tr>
<tr>
<td>X=“bird”</td>
<td>.1</td>
<td>.1</td>
</tr>
<tr>
<td>X=“human”</td>
<td>.1</td>
<td>.1</td>
</tr>
</tbody>
</table>

\begin{center}
\text{Joint probability}
\end{center}
Joint Probabilities

what happens as we add conjuncts?
Joint Probabilities

what happens as we add conjuncts?
Joint Probabilities

what happens as we add conjuncts?
Joint Probabilities

Joint probabilities: $p(A, B, C, D)$

$p(A, B, C)$

$p(A, B)$

$p(A)$

$\text{what happens as we add conjuncts?}$
Joint Probabilities

what happens as we add conjuncts?
## Probability Prerequisites

<table>
<thead>
<tr>
<th>Basic probability axioms and definitions</th>
<th>Bayes rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joint probability</td>
<td>Probability chain rule</td>
</tr>
<tr>
<td>Probabilistic Independence</td>
<td>Common distributions</td>
</tr>
<tr>
<td>Marginal probability</td>
<td>Expected Value (of a function) of a Random Variable</td>
</tr>
<tr>
<td>Definition of conditional probability</td>
<td></td>
</tr>
</tbody>
</table>
Probabilistic Independence

Independence: when events can occur and not impact the probability of other events

Formally: $p(x,y) = p(x) \cdot p(y)$

Generalizable to $> 2$ random variables

Q: Are the results of flipping the same coin twice in succession independent?
Probabilistic Independence

Independence: when events can occur and not impact the probability of other events

Formally: $p(x,y) = p(x) \times p(y)$

Generalizable to $>2$ random variables

Q: Are the results of flipping the same coin twice in succession independent?

A: Yes (assuming no weird effects)
Probabilistic Independence

Independence: when events can occur and not impact the probability of other events

Formally: \( p(x,y) = p(x) \times p(y) \)

Generalizable to > 2 random variables

Q: Are A and B independent?
Probabilistic Independence

Independence: when events can occur and not impact the probability of other events

Formally: \( p(x,y) = p(x) \cdot p(y) \)

Generalizable to > 2 random variables

Q: Are A and B independent?

A: No (work it out from \( p(A,B) \)) and the axioms
Probabilistic Independence

Independence: when events can occur and not impact the probability of other events

Formally: \( p(x,y) = p(x) \times p(y) \)

Generalizable to > 2 random variables

Q: Are X and Y independent?

<table>
<thead>
<tr>
<th>( p(x,y) )</th>
<th>Y=0</th>
<th>Y=1</th>
</tr>
</thead>
<tbody>
<tr>
<td>X=&quot;cat&quot;</td>
<td>.04</td>
<td>.32</td>
</tr>
<tr>
<td>X=&quot;dog&quot;</td>
<td>.2</td>
<td>.04</td>
</tr>
<tr>
<td>X=&quot;bird&quot;</td>
<td>.1</td>
<td>.1</td>
</tr>
<tr>
<td>X=&quot;human&quot;</td>
<td>.1</td>
<td>.1</td>
</tr>
</tbody>
</table>
Probabilistic Independence

Independence: when events can occur and not impact the probability of other events

Formally: \( p(x,y) = p(x) \times p(y) \)

Generalizable to > 2 random variables

Q: Are \( X \) and \( Y \) independent?

<table>
<thead>
<tr>
<th>( p(x,y) )</th>
<th>( Y=0 )</th>
<th>( Y=1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X=\text{“cat”} )</td>
<td>.04</td>
<td>.32</td>
</tr>
<tr>
<td>( X=\text{“dog”} )</td>
<td>.2</td>
<td>.04</td>
</tr>
<tr>
<td>( X=\text{“bird”} )</td>
<td>.1</td>
<td>.1</td>
</tr>
<tr>
<td>( X=\text{“human”} )</td>
<td>.1</td>
<td>.1</td>
</tr>
</tbody>
</table>

A: No (find the marginal probabilities of \( p(x) \) and \( p(y) \))
Probability Prerequisites

- Basic probability axioms and definitions
- Joint probability
- Probabilistic Independence
- Marginal probability
- Definition of conditional probability
- Bayes rule
- Probability chain rule
- Common distributions
- Expected Value (of a function) of a Random Variable
Marginal(ized) Probability: The Discrete Case

Consider the mutually exclusive ways that different values of $x$ could occur with $y$

Q: How do write this in terms of joint probabilities?
Marginal(ized) Probability: The Discrete Case

Consider the **mutually exclusive** ways that different values of $x$ could occur with $y$

$$p(y) = \sum_x p(x, y)$$
Probability Prerequisites

Basic probability axioms and definitions

Joint probability

Probabilistic Independence

Marginal probability

Definition of conditional probability

Bayes rule

Probability chain rule

Common distributions

Expected Value (of a function) of a Random Variable
Conditional Probability

\[ p(X \mid Y) = \frac{p(X, Y)}{p(Y)} \]
Conditional Probability

\[ p(X \mid Y) = \frac{p(X, Y)}{p(Y)} \]

\[ p(Y) = \text{marginal probability of } Y \]
Conditional Probability

\[ p(X \mid Y) = \frac{p(X, Y)}{p(Y)} \]

\[ p(Y) = \int p(X, Y) \, dX \]
Conditional Probabilities: Changing the Right

what happens as we add conjuncts to the right?
Conditional Probabilities: Changing the Right

\[ p(A) \]
\[ p(A \mid B) \]

what happens as we add conjuncts to the right?
Conditional Probabilities: Changing the Right

what happens as we add conjuncts to the right?
Conditional Probabilities: Changing the Right

what happens as we add conjuncts to the right?
Conditional Probabilities

Bias vs. Variance

Lower bias: More specific to what we care about

Higher variance: For fixed observations, estimates become less reliable
Revisiting Marginal Probability: The Discrete Case

\[ p(y) = \sum_x p(x, y) \]

\[ = \sum_x p(x)p(y \mid x) \]
Probability Prerequisites

Basic probability axioms and definitions

Joint probability

Probabilistic Independence

Marginal probability

Definition of conditional probability

Bayes rule

Probability chain rule

Common distributions

Expected Value (of a function) of a Random Variable
Deriving Bayes Rule

Start with conditional

\[ p(X \mid Y) \]
Deriving Bayes Rule

\[ p(X \mid Y) = \frac{p(X, Y)}{p(Y)} \]
Deriving Bayes Rule

\[ p(X \mid Y) = \frac{p(X, Y)}{p(Y)} \]

\[ p(X, Y) = p(X \mid Y)p(Y) \]

\[ p(X \mid Y) = \frac{p(Y \mid X) \cdot p(X)}{p(Y)} \]
Bayes Rule

\[ p(X \mid Y) = \frac{p(Y \mid X) \times p(X)}{p(Y)} \]

- posterior probability
- likelihood
- prior probability
- marginal likelihood
- (probability)
Probability Prerequisites

- Basic probability axioms and definitions
- Joint probability
- Probabilistic Independence
- Marginal probability
- Definition of conditional probability
- Bayes rule
- Probability chain rule
- Common distributions
- Expected Value (of a function) of a Random Variable
Probability Chain Rule

\[ p(x_1, x_2, \ldots, x_S) = \prod_{i=1}^{S} p(x_i | x_1, \ldots, x_{i-1}) \]

extension of Bayes rule
Probability Prerequisites

Basic probability axioms and definitions

Joint probability

Probabilistic Independence

Marginal probability

Definition of conditional probability

Bayes rule

Probability chain rule

Common distributions

Expected Value (of a function) of a Random Variable
Distribution Notation

If $X$ is a R.V. and $G$ is a distribution:

- $X \sim G$ means $X$ is distributed according to ($\text{"sampled from"}$) $G$
Distribution Notation

If X is a R.V. and G is a distribution:

• \( X \sim G \) means X is distributed according to ("sampled from") \( G \)

• \( G \) often has parameters \( \rho = (\rho_1, \rho_2, \ldots, \rho_M) \) that govern its "shape"

• Formally written as \( X \sim G(\rho) \)
Distribution Notation

If X is a R.V. and G is a distribution:

• $X \sim G$ means X is distributed according to ("sampled from") $G$
• $G$ often has parameters $\rho = (\rho_1, \rho_2, \ldots, \rho_M)$ that govern its "shape"
• Formally written as $X \sim G(\rho)$

**i.i.d.** If $X_1, X_2, \ldots, X_N$ are all independently sampled from $G(\rho)$, they are independently and identically distributed
Common Distributions

Bernoulli/Binomial
- Binary R.V.: 0 (failure) or 1 (success)
- $X \sim \text{Bernoulli}(\rho)$
- $p(X = 1) = \rho$, $p(X = 0) = 1 - \rho$
- Generally, $p(X = k) = \rho^k (1 - \rho)^{1-k}$

Categorical/Multinomial

Poisson

Normal

(Gamma)
Common Distributions

Bernoulli/Binomial

- Bernoulli: A single draw
  - Binary R.V.: 0 (failure) or 1 (success)
  - $X \sim \text{Bernoulli}(\rho)$
  - $p(X = 1) = \rho, p(X = 0) = 1 - \rho$
  - Generally, $p(X = k) = \rho^k (1 - \rho)^{1-k}$

Binomial: Sum of N iid Bernoulli draws

- Values $X$ can take: 0, 1, ..., N
- Represents number of successes
- $X \sim \text{Binomial}(N, \rho)$
- $p(X = k) = \binom{N}{k} \rho^k (1 - \rho)^{N-k}$
Common Distributions

Bernoulli/Binomial

Categorical/Multinomial

Poisson

Normal

(Gamma)

Categorical: A single draw
- Finite R.V. taking one of K values: 1, 2, ..., K
- \( X \sim \text{Cat}(\rho), \rho \in \mathbb{R}^K \)
- \( p(X = 1) = \rho_1, p(X = 2) = \rho_2, \ldots p(X = K) = \rho_K \)
- Generally, \( p(X = k) = \prod_j \rho_j^{1[k=j]} \)
- \( 1[c] = \begin{cases} 1, & c \text{ is true} \\ 0, & c \text{ is false} \end{cases} \)

Multinomial: Sum of N iid Categorical draws
- Vector of size K representing how often value k was drawn
- \( X \sim \text{Multinomial}(N, \rho), \rho \in \mathbb{R}^K \)
Common Distributions

Poisson

- Finite R.V. taking any integer that is $\geq 0$
- $X \sim \text{Poisson}(\lambda), \lambda \in \mathbb{R}$ is the "rate"
- $p(X = k) = \frac{\lambda^k \exp(-\lambda)}{k!}$
Common Distributions

Normal

- Real R.V. taking any real number
- $X \sim \text{Normal}(\mu, \sigma)$, $\mu$ is the mean, $\sigma$ is the standard deviation

\[
p(X = x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)
\]
Probability Prerequisites

- Basic probability axioms and definitions
- Joint probability
- Probabilistic Independence
- Marginal probability
- Definition of conditional probability
- Bayes rule
- Probability chain rule
- Common distributions
- Expected Value (of a function) of a Random Variable
Expected Value of a Random Variable

\[ X \sim p(\cdot) \]
Expected Value of a Random Variable

\[ X \sim p(\cdot) \]

\[ \mathbb{E}[X] = \sum_x x \, p(x) \]
Expected Value: Example

uniform distribution of number of cats I have

\[ \mathbb{E}[X] = \sum_{x} x \cdot p(x) \]

\[ = \frac{1}{6} \times 1 + \frac{1}{6} \times 2 + \frac{1}{6} \times 3 + \frac{1}{6} \times 4 + \frac{1}{6} \times 5 + \frac{1}{6} \times 6 \]

= 3.5
Expected Value: Example

uniform distribution of number of cats I have

$$\mathbb{E}[X] = \sum_x x \, p(x)$$

$$\frac{1}{6} \times 1 + \frac{1}{6} \times 2 + \frac{1}{6} \times 3 + \frac{1}{6} \times 4 + \frac{1}{6} \times 5 + \frac{1}{6} \times 6 = 3.5$$

Q: What common distribution is this?
Expected Value: Example

uniform distribution of number of cats I have

\[ E[X] = \sum_{x} x p(x) \]

\[ = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 \]

\[ = 3.5 \]

Q: What common distribution is this?

A: Categorical
Expected Value: Example 2

non-uniform distribution of number of cats a normal cat person has

\[ E[X] = \sum_{x} x p(x) = \frac{1}{2} \cdot 1 + \frac{1}{10} \cdot 2 + \frac{1}{10} \cdot 3 + \frac{1}{10} \cdot 4 + \frac{1}{10} \cdot 5 + \frac{1}{10} \cdot 6 = 2.5 \]
Expected Value of a Function of a Random Variable

\[ X \sim p(\cdot) \]

\[ \mathbb{E}[X] = \sum_x x \, p(x) \]

\[ \mathbb{E}[f(X)] = ??? \]
Expected Value of a Function of a Random Variable

\[ X \sim p(\cdot) \]

\[ \mathbb{E}[X] = \sum_{x} x \ p(x) \]

\[ \mathbb{E}[f(X)] = \sum_{x} f(x) \ p(x) \]
Expected Value of Function: Example

non-uniform distribution of number of cats I start with

What if each cat magically becomes two?

\[ f(k) = 2^k \]

\[ \mathbb{E}[f(X)] = \sum_x f(x) p(x) \]
Expected Value of Function: Example

non-uniform distribution of number of cats I start with

What if each cat magically becomes two?

\[ f(k) = 2^k \]

\[ \mathbb{E}[f(X)] = \sum_{x} f(x) p(x) = \sum_{x} 2^x p(x) \]

\[ \frac{1}{2} \times 2^1 + \frac{1}{10} \times 2^2 + \frac{1}{10} \times 2^3 + \frac{1}{10} \times 2^4 + \frac{1}{10} \times 2^5 + \frac{1}{10} \times 2^6 = 13.4 \]
Probability Prerequisites

Basic probability axioms and definitions

Joint probability

Probabilistic Independence

Marginal probability

Definition of conditional probability

Bayes rule

Probability chain rule

Common distributions

Expected Value (of a function) of a Random Variable
Outline

Review+Extension

Probability

Decision Theory

Loss Functions
Decision Theory

“Decision theory is trivial, apart from the computational details” – MacKay, ITILA, Ch 36

Input: \( x \) ("state of the world")

Output: a decision \( \hat{y} \)
Decision Theory

“Decision theory is trivial, apart from the computational details” – MacKay, ITILA, Ch 36

Input: \( x \) (“state of the world”)
Output: a decision \( \hat{y} \)
Requirement 1: a decision (hypothesis) function \( h(x) \) to produce \( \hat{y} \)
Decision Theory

“Decision theory is trivial, apart from the computational details” – MacKay, ITILA, Ch 36

Input: $\mathbf{x}$ (“state of the world”)
Output: a decision $\hat{y}$

Requirement 1: a decision (hypothesis) function $h(\mathbf{x})$ to produce $\hat{y}$

Requirement 2: a function $\ell(y, \hat{y})$ telling us how wrong we are
Decision Theory

“Decision theory is trivial, apart from the computational details” – MacKay, ITILA, Ch 36

Input: \( x \) (“state of the world”)
Output: a decision \( \hat{y} \)

Requirement 1: a decision (hypothesis) function \( h(x) \) to produce \( \hat{y} \)

Requirement 2: a loss function \( \ell(y, \hat{y}) \) telling us how wrong we are

Goal: minimize our expected loss across any possible input
Requirement 1: Decision Function

$h(x)$ is our predictor (classifier, regression model, clustering model, etc.)
Requirement 2: Loss Function

\[ \ell(y, \hat{y}) \geq 0 \]

loss: A function that tells you how much to penalize a prediction \( \hat{y} \) from the correct answer \( y \)

\( \ell \) ("ell" fancy l character)  

predicted label/result

"correct" label/result

optimize \( \ell \)?  
- minimize
- maximize
Requirement 2: Loss Function

A loss function, $\ell (y, \hat{y}) \geq 0$, is a function that tells you how much to penalize a prediction $\hat{y}$ from the correct answer $y$. Negative $\ell (-\ell)$ is called a utility or reward function.

- $y$ represents the "correct" label/result.
- $\hat{y}$ represents the predicted label/result.

The "ell" (fancy l character) notation is used to denote the loss function.
Decision Theory

minimize expected loss across any possible input

$$\arg \min_{\hat{y}} \mathbb{E}[\ell(y, \hat{y})]$$
Risk Minimization

minimize expected loss across any possible input

\[ \arg \min_{\hat{y}} \mathbb{E}[\ell(y, \hat{y})] = \arg \min_{h} \mathbb{E}[\ell(y, h(x))] \]

a particular, unspecified input pair \((x, y)\) ... but we want any possible pair
Decision Theory

minimize expected loss across any possible input

\[
\arg \min \mathbb{E}[\ell(y, \hat{y})] = \\
\arg \min \mathbb{E}[\ell(y, h(x))] = \\
\arg\min_h \mathbb{E}_{(x, y) \sim P} [\ell(y, h(x))]
\]

Assumption: there exists some true (but likely unknown) distribution \(P\) over inputs \(x\) and outputs \(y\)
Risk Minimization

minimize expected loss across any possible input

\[
\arg \min_{\hat{y}} \mathbb{E}[\ell(y, \hat{y})] = \\
\arg \min_h \mathbb{E}[\ell(y, h(x))] = \\
\arg\min_h \mathbb{E}_{(x,y) \sim P}[\ell(y, h(x))] = \\
\arg\min_h \int \ell(y, h(x)) P(x, y) d(x, y)
\]
Risk Minimization

minimize expected loss across any possible input

$$\arg \min_{\hat{y}} \mathbb{E}[\ell(y, \hat{y})] =$$

$$\arg \min_h \mathbb{E}[\ell(y, h(x))] =$$

$$\arg\!\min_h \mathbb{E}_{(x,y) \sim P}[\ell(y, h(x))] =$$

$$\arg\!\min_h \int \ell(y, h(x)) P(x, y) d(x, y)$$

*we don’t know this distribution*! 

*we could try to approximate it analytically*
Empirical Risk Minimization

minimize expected loss across our observed input

$$\arg\min_{\hat{y}} \mathbb{E}[\ell(y, \hat{y})] =$$

$$\arg\min_{h} \mathbb{E}[\ell(y, h(x))] =$$

$$\arg\min_{h} \mathbb{E}_{(x,y) \sim P} \left[ \ell(y, h(x)) \right] \approx$$

$$\arg\min_{h} \frac{1}{N} \sum_{i=1}^{N} \ell(y_i, h(x_i))$$
Empirical Risk Minimization

minimize expected loss across our observed input

$$\arg\min_h \sum_{i=1}^N \ell(y_i, h(x_i))$$

our classifier/predictor controlled by our parameters $\theta$

change $\theta \rightarrow$ change the behavior of the classifier
Best Case: Optimize Empirical Risk with Gradients

$$\arg\min_h \sum_{i=1}^{N} \ell(y_i, h_\theta(x_i))$$

change $\theta \rightarrow$
change the behavior of the classifier

$$\arg\min_\theta \sum_{i=1}^{N} \ell(y_i, h_\theta(x_i))$$
Best Case: Optimize Empirical Risk with Gradients

\[
\arg\min_{\theta} \sum_{i=1}^{N} \ell(y_i, h_{\theta}(x_i))
\]

How? Use Gradient Descent on \( F(\theta) \)!

*differentiating might not always work: “... apart from the computational details”*
Best Case: Optimize Empirical Risk with Gradients

\[
\arg\min_{\theta} \sum_{i=1}^{N} \ell(y_i, h_\theta(x_i))
\]

change \( \theta \rightarrow \)
change the behavior of the classifier

\[
\nabla_\theta F = \sum_i \frac{\partial \ell(y_i, \hat{y} = h_\theta(x_i))}{\partial \hat{y}} \nabla_\theta h_\theta(x_i)
\]

differentiating might not always work: “... apart from the computational details”
Best Case: Optimize Empirical Risk with Gradients

$$\arg\min_{\theta} \sum_{i=1}^{N} \ell(y_i, h_{\theta}(x_i))$$

Change $\theta \rightarrow$ change the behavior of the classifier

$$\nabla_{\theta} F = \sum_{i} \frac{\partial \ell(y_i, \hat{y} = h_{\theta}(x_i))}{\partial \hat{y}} \nabla_{\theta} h_{\theta}(x_i)$$

*Step 1: compute the gradient of the loss wrt the predicted value*

differentiating might not always work: “… apart from the computational details”
Best Case: Optimize Empirical Risk with Gradients

\[
\arg\min_\theta \sum_{i=1}^N \ell(y_i, h_\theta(x_i))
\]

change $\theta \rightarrow$
change the behavior of the classifier

\[
\nabla_\theta F = \sum_i \frac{\partial \ell(y_i, \hat{y} = h_\theta(x_i))}{\partial \hat{y}} \nabla_\theta h_\theta(x_i)
\]

Step 1: compute the gradient of the loss wrt the predicted value

Step 2: compute the gradient of the predicted value wrt $\theta$.

differentiating might not always work: “... apart from the computational details”
Outline

Review+Extension

Probability

Decision Theory

Loss Functions
Loss Functions Serve a Task

Classification
Regression
Clustering

the task: what kind of problem are you solving?

Fully-supervised
Semi-supervised
Un-supervised

the data: amount of human input/number of labeled examples

Probabilistic
Generative
Conditional
Spectral
Neural
Memory-based
Exemplar
...

the approach: how any data are being used
Classification: Supervised Machine Learning

Assigning subject categories, topics, or genres
Spam detection
Authorship identification

Input:
- an instance \( d \)
- a fixed set of classes \( C = \{c_1, c_2, \ldots, c_J\} \)
- A training set of \( m \) hand-labeled instances \((d_1, c_1), \ldots, (d_m, c_m)\)

Output:
- a learned classifier \( \gamma \) that maps instances to classes

\( \gamma \) learns to associate certain \textit{features} of instances with their labels
### Classification Example: Face Recognition

<table>
<thead>
<tr>
<th>Class</th>
<th>Image</th>
<th>Class</th>
<th>Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avrim</td>
<td><img src="image1.jpg" alt="Avrim Image 1" /></td>
<td>Tom</td>
<td><img src="image2.jpg" alt="Tom Image 1" /></td>
</tr>
<tr>
<td>Avrim</td>
<td><img src="image1.jpg" alt="Avrim Image 2" /></td>
<td>Tom</td>
<td><img src="image2.jpg" alt="Tom Image 2" /></td>
</tr>
<tr>
<td>Avrim</td>
<td><img src="image1.jpg" alt="Avrim Image 3" /></td>
<td>Tom</td>
<td><img src="image2.jpg" alt="Tom Image 3" /></td>
</tr>
<tr>
<td>Avrim</td>
<td><img src="image1.jpg" alt="Avrim Image 4" /></td>
<td>Tom</td>
<td><img src="image2.jpg" alt="Tom Image 4" /></td>
</tr>
<tr>
<td>Avrim</td>
<td><img src="image1.jpg" alt="Avrim Image 5" /></td>
<td>Tom</td>
<td><img src="image2.jpg" alt="Tom Image 5" /></td>
</tr>
</tbody>
</table>

Courtesy from Hamed Pirsiavash
Classification Loss Function Example:
0-1 Loss

\[ \ell(y, \hat{y}) = \begin{cases} 
0, & \text{if } y = \hat{y} \\
1, & \text{if } y \neq \hat{y} 
\end{cases} \]
Classification Loss Function Example: 0-1 Loss

\[ \ell (y, \hat{y}) = \begin{cases} 
0, & \text{if } y = \hat{y} \\
1, & \text{if } y \neq \hat{y} 
\end{cases} \]

Problem 1: not differentiable wrt \( \hat{y} \) (or \( \theta \))
Classification Loss Function Example: 0-1 Loss

\[ \ell(y, \hat{y}) = \begin{cases} 
0, & \text{if } y = \hat{y} \\
1, & \text{if } y \neq \hat{y} 
\end{cases} \]

Problem 1: not differentiable wrt \( \hat{y} \) (or \( \theta \))

Solution 1: is the data linearly separable? Perceptron (next class) can work
Classification Loss Function Example: 0-1 Loss

\[
\ell(y, \hat{y}) = \begin{cases} 
0, & \text{if } y = \hat{y} \\
1, & \text{if } y \neq \hat{y}
\end{cases}
\]

Problem 1: not differentiable wrt \( \hat{y} \) (or \( \theta \))

Solution 1: is the data linearly separable? Perceptron (next class) can work

Solution 2: is \( h(x) \) a conditional distribution \( p(y \mid x) \)? Maximize that probability (a couple classes)
Structured Classification: Sequence & Structured Prediction

---

- **Sentence:**
  - Noun phrase
  - Determiner
  - Noun
  - Verb phrase
  - Determiner
  - Adjective
  - Noun

- **Example:**
  - The dog chased the black cat.

---

- **Translated Text:**

  ऑस्ट्रेलिया में खेली जा रही त्रि-सीरीज क्रिकेट सीरीज के दौरान का एक दिनांक एक वर्क्ष क्रिकेट मैच का यह सुरुवात में साध्य हो सकता है।

  भारत इस सीरीज के साथ ऑस्ट्रेलिया के लिए एक वर्क्ष क्रिकेट मैच की सीरीज में 0-2 से हार गया था।

  अन्तरराष्ट्रीय क्रिकेट के साथ मुस्लिम जनता के बाद भारत के कप्तान के संघर्ष ने तीनों शहीद भी तीनों ने टेस्ट क्रिकेट के दौरान का एक दिनांक का यह सुरुवात में साध्य हो सकता है।

---

- **Images:**

  - Cows grazing
  - People walking
  - Cartoon animals and objects

---

**Courtesy Hamed Pirsiavash**
Structured Classification Loss Function

Example: 0-1 Loss?

\[ \ell(y, \hat{y}) = \begin{cases} 
0, & \text{if } y = \hat{y} \\
1, & \text{if } y \neq \hat{y} 
\end{cases} \]

Problem 1: not differentiable wrt \( \hat{y} \) (or \( \theta \))

Solution 1: is the data linearly separable? Perceptron (next class) can work

Solution 2: is \( h(x) \) a conditional distribution \( p(y \mid x) \)? Use MAP

Problem 2: too strict. Structured Prediction involves many individual decisions

Solution 1: Specialize 0-1 to the structured problem at hand
Regression

Like classification, but real-valued
Regression Example: Stock Market Prediction

S&P 500
S&P Indices: .INX - Jan 16 4:30 PM ET

2,019.42 ↑26.75 (1.34%)

1 day | 5 day | 1 month | 3 month | 1 year | 5 year | max

Open 1,992.25
High 2,020.46
Low 1,988.12

Market cap -
P/E ratio (ttm) -
Dividend yield -

Previous close 1,992.67

Courtesy Hamed Pirsiavash
Regression Loss Function Examples

squared loss/MSE (Mean squared error)

\[ \ell(y, \hat{y}) = (y - \hat{y})^2 \]

\(\hat{y}\) is a real value \(\rightarrow\) nicely differentiable (generally) 😊
Regression Loss Function Examples

Squared loss/MSE (Mean squared error)

\[
\ell(y, \hat{y}) = (y - \hat{y})^2
\]

\(\hat{y}\) is a real value \(\rightarrow\) nicely differentiable (generally) 😊

Absolute loss

\[
\ell(y, \hat{y}) = |y - \hat{y}|
\]

Absolute value is mostly differentiable
Regression Loss Function Examples

squared loss/MSE (Mean squared error)

\[ \ell(y, \hat{y}) = (y - \hat{y})^2 \]

\(\hat{y}\) is a real value \(\rightarrow\) nicely differentiable (generally) 😊

absolute loss

\[ \ell(y, \hat{y}) = |y - \hat{y}| \]

Absolute value is *mostly* differentiable

These loss functions prefer different behavior in the predictions (hint: look at the gradient of each)... we’ll get back to this
Unsupervised learning: Clustering

We’ll return to clustering loss functions later

Courtesy Hamed Pirsiavash
Outline

Review+Extension

Probability

Decision Theory

Loss Functions