ECE 5424: Introduction to Machine Learning

Topics:
- Probability Review

Readings: Barber 8.1, 8.2

Stefan Lee
Virginia Tech
Project

• Groups of 1-3
  – we prefer teams of 2

• Deliverables:
  – Project proposal (NIPS format): 2 page, due Sept 21
  – Midway presentations (in class)
  – Final report: webpage with results
Administrative

• HW1
  – Due on Wed 09/14, 11:55pm

• Project Proposal
  – Due: Wed 09/21, 11:55 pm
  – <=2pages, NIPS format
Proposal

• 2 Page (NIPS format)
  – https://nips.cc/Conferences/2015/PaperInformation/StyleFiles

• Necessary Information:
  – Project title
  – Project idea.
    • This should be approximately two paragraphs.
  – Data set details
    • Ideally existing dataset. No data-collection projects.
  – Software
    • Which libraries will you use?
    • What will you write?
  – Papers to read.
    • Include 1-3 relevant papers. You will probably want to read at least one of them before submitting your proposal.
  – Teammate
    • Will you have a teammate? If so, what’s the break-down of labor? Maximum team size is 3 students.
  – Mid-semester Milestone
    • What will you complete by the project milestone due date? Experimental results of some kind are expected here.
Project

• Rules
  – Must be about machine learning
  – Must involve real data
    • Use your own data or take from class website
  – Can apply ML to your own research.
    • Must be done this semester.
  – OK to combine with other class-projects
    • Must declare to both course instructors
    • Must have explicit permission from BOTH instructors
    • Must have a sufficient ML component
  – Using libraries
    • No need to implement all algorithms
    • OK to use standard SVM, MRF, Decision-Trees, etc libraries
    • More thought + effort => More credit
Project

• **Main categories**
  – Application/Survey
    • Compare a bunch of existing algorithms on a new application domain of your interest
  – Formulation/Development
    • Formulate a new model or algorithm for a new or old problem
  – Theory
    • Theoretically analyze an existing algorithm

• **Support**
  – List of ideas, pointers to dataset/algorithms/code
    • [https://filebox.ece.vt.edu/~f16ece5424/project.html](https://filebox.ece.vt.edu/~f16ece5424/project.html)
    • We will mentor teams and give feedback.
Procedural View

• Training Stage:
  – Raw Data $\rightarrow x$ (Feature Extraction)
  – Training Data $\{ (x,y) \} \rightarrow f$ (Learning)

• Testing Stage
  – Raw Data $\rightarrow x$ (Feature Extraction)
  – Test Data $x \rightarrow f(x)$ (Apply function, Evaluate error)
Statistical Estimation View

• Probabilities to rescue:
  – $x$ and $y$ are random variables
  – $D = (x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N) \sim P(X, Y)$

• IID: Independent Identically Distributed
  – Both training & testing data sampled IID from $P(X, Y)$
  – Learn on training set
  – Have some hope of *generalizing* to test set
Plan for Today

• Review of Probability
  – Discrete vs Continuous Random Variables
  – PMFs vs PDF
  – Joint vs Marginal vs Conditional Distributions
  – Bayes Rule and Prior
  – Expectation, Entropy, KL-Divergence
Probability

• The world is a very uncertain place

• 30 years of Artificial Intelligence and Database research danced around this fact

• And then a few AI researchers decided to use some ideas from the eighteenth century
Probability

• A is non-deterministic event
  – Can think of A as a boolean-valued variable

• Examples
  – A = your next patient has cancer
  – A = Donald Trump Wins the 2016 Presidential Election
Interpreting Probabilities

• What does $P(A)$ mean?

• Frequentist View
  – limit $N \to \infty \frac{\#(A \text{ is true})}{N}$
  – limiting frequency of a repeating non-deterministic event

• Bayesian View
  – $P(A)$ is your “belief” about $A$

• Market Design View
  – $P(A)$ tells you how much you would bet
The Axioms Of Probability
Axioms of Probability

- $0 \leq P(A) \leq 1$
- $P($empty-set$) = 0$
- $P($everything$) = 1$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
Interpreting the Axioms

• $0 \leq P(A) \leq 1$
• $P(\text{empty-set}) = 0$
• $P(\text{everything}) = 1$
• $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

Event space of all possible worlds

Its area is 1

Worlds in which $A$ is True

$P(A) = \text{Area of reddish oval}$

Worlds in which $A$ is False
Interpreting the Axioms

• $0 \leq P(A) \leq 1$
• $P(\text{empty-set}) = 0$
• $P(\text{everything}) = 1$
• $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

The area of A can’t get any smaller than 0

And a zero area would mean no world could ever have A true

(C) Dhruv Batra

Image Credit: Andrew Moore
Interpreting the Axioms

- $0 \leq P(A) \leq 1$
- $P(\text{empty-set}) = 0$
- $P(\text{everything}) = 1$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

The area of $A$ can’t get any bigger than 1

And an area of 1 would mean all worlds will have A true
Interpreting the Axioms

• $0 \leq P(A) \leq 1$
• $P(\text{empty-set}) = 0$
• $P(\text{everything}) = 1$
• $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

Simple addition and subtraction
Concepts

• Sample Space
  – Space of events

• Random Variables
  – Mapping from events to numbers
  – Discrete vs Continuous

• Probability
  – Mass vs Density
**Discrete Random Variables**

\[ X \quad \text{discrete random variable} \]

\[ \mathcal{X} \quad \text{or } \text{Val}(X) \quad \text{sample space of possible outcomes, which may be finite or countably infinite} \]

\[ x \in \mathcal{X} \quad \text{outcome of sample of discrete random variable} \]

\[ p(X = x) \quad \text{probability distribution (probability mass function)} \]

\[ p(x) \quad \text{shorthand used when no ambiguity} \]

\[ 0 \leq p(x) \leq 1 \quad \text{for all } x \in \mathcal{X} \]

\[ \sum_{x \in \mathcal{X}} p(x) = 1 \]

\[ \mathcal{X} = \{1, 2, 3, 4\} \quad \text{uniform distribution} \]

\[ \text{degenerate distribution} \]
Continuous Random Variables

• On board
Concepts

• Expectation

• Variance
Most Important Concepts

• Marginal distributions / Marginalization

• Conditional distribution / Chain Rule

• Bayes Rule
Joint Distribution
Marginalization

• Marginalization
  – Events: \( P(A) = P(A \text{ and } B) + P(A \text{ and } \text{not } B) \)
  
  – Random variables \( P(X = x) = \sum_y P(X = x, Y = y) \)
Marginal Distributions

\[ p(x, y) = \sum_{z \in Z} p(x, y, z) \]

\[ p(x) = \sum_{y \in Y} p(x, y) \]
Conditional Probabilities

- $P(Y=y \mid X=x)$

- What do you believe about $Y=y$, if I tell you $X=x$?

- $P(\text{Donald Trump Wins the 2016 Election})$?

- What if I tell you:
  - He has the Republican nomination
  - His twitter history
  - The complete DVD set of The Apprentice
Conditional Probabilities

• $P(A \mid B) = \text{In worlds that where B is true, fraction where A is true}$

• Example
  - H: “Have a headache”
  - F: “Coming down with Flu”

  \[
P(H) = \frac{1}{10} \\
P(F) = \frac{1}{40} \\
P(H\mid F) = \frac{1}{2}
\]

  “Headaches are rare and flu is rarer, but if you’re coming down with ‘flu there’s a 50-50 chance you’ll have a headache.”
Conditional Distributions

\[ p(x, y \mid Z = z) = \frac{p(x, y, z)}{p(z)} \]
Conditional Probabilities

• Definition

• Corollary: Chain Rule
Independent Random Variables

\[ P(x,y) = p(x)p(y) \text{ for all } x \in \mathcal{X}, y \in \mathcal{Y} \]
Marginal Independence

- **Sets** of variables $X, Y$

- $X$ is independent of $Y$
  - Shorthand: $P \vdash (X \perp Y)$

- **Proposition:** $P$ satisfies $(X \perp Y)$ if and only if
  - $P(X=x, Y=y) = P(X=x) P(Y=y)$, $\forall x \in Val(X), \forall y \in Val(Y)$
Conditional independence

• **Sets** of variables $X, Y, Z$

• $X$ is independent of $Y$ given $Z$ if
  - Shorthand: $P \vdash (X \perp Y \mid Z)$
  - For $P \vdash (X \perp Y \mid Z)$, write $P \vdash (X \perp Y)$

• **Proposition:** $P$ satisfies $(X \perp Y \mid Z)$ if and only if
  - $P(X,Y|Z) = P(X|Z) \cdot P(Y|Z), \quad \forall x \in Val(X), \forall y \in Val(Y), \forall z \in Val(Z)$
Concept

• Bayes Rules
  – Simple yet fundamental

\[
P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A|B) P(B)}{P(A)}
\]

This is Bayes Rule

Bayes Rule

• Simple yet profound
  – Using Bayes Rules doesn’t make your analysis Bayesian!

• Concepts:
  – Likelihood
    • How much does a certain hypothesis explain the data?
  – Prior
    • What do you believe before seeing any data?
  – Posterior
    • What do we believe after seeing the data?
Entropy

- Measures the amount of ambiguity or uncertainty in a distribution:

\[ H(p) = - \sum_x p(x) \log p(x) \]

- Expected value of \(-\log p(x)\) (a function which depends on \(p(x)!\)).
- \(H(p) > 0\) unless only one possible outcome in which case \(H(p) = 0\).
- Maximal value when \(p\) is uniform.
- Tells you the expected "cost" if each event costs \(-\log p(\text{event})\)
KL-Divergence / Relative Entropy

• An asymmetric measure of the distance between two distributions:

\[
KL[p\|q] = \sum_x p(x) \log \frac{p(x)}{q(x)}
\]

• \( KL > 0 \) unless \( p = q \) then \( KL = 0 \)

• Tells you the extra cost if events were generated by \( p(x) \) but instead of charging under \( p(x) \) you charged under \( q(x) \).
$p(x) \quad q(x)$

Original Gaussian PDF's

$D_{KL} (P \| Q)$

KL Area to be Integrated

\( p(x) \) \( q(x) \)

Original Gaussian PDF's

KL Area to be Integrated

\( D_{KL}(P \parallel Q) \)

\( D_{KL}(P \parallel Q) \)
• End of Prob. Review

• Start of Estimation