Note

proj_topic - due - Oct 18

• Homework #2 submission closed.
• Start to work on choosing final project topic.
  – Read LeCun et al. if not done so yet.
  – Pick a subtopic in the paper that you like to go deeper.
  – Select main papers among those cited in the section describing your topic in LeCun et al.
  – Continue to select papers cited among the cited papers.
  – Finalize at least 3 papers on your subtopic.
  – Submit your subtopic/summary/selected papers to iLearn forum for my review (make your own thread).
  – Due on Oct 18

PF
Framework for Bayesian Probabilistic Reasoning

CSC 872
Pattern Analysis and Machine Intelligence
### Logical Reasoning: Review

- Logical languages (PL, FOL) gave us formal ways to describe a world based on **Boolean truth**
  - World with *True/False*
  - PL’s ontological commitment: **Facts**
  - FOL’s ontological commitment: **Facts, Objects, Relations**
- **Inference**: *derive a new fact from known facts (KB) !!!*
  - Sound & Complete Inference = Entailment
- **BUT** we live in a world that is very uncertain!
  - The car accident example from the last lecture / tomorrow’s weather / stock market / your and my life even! …
- **How to represent facts with such uncertainty?**
- **How to do inference with such uncertainty?**

### Probabilistic Reasoning

- One answer is …
- **Bayesian Probability!**
  - OC: Proposition (Facts)
  - OC: **Degree of Belief**
  - Inference by **Bayes Rule**
- **Frequentist Approach**
  - Define a prob. as a limit of event’s frequency in a large number of trials
  - More objective it seems: used for hypothesis testing … **BUT**
  - Probability that “it will rain tomorrow”?
  - It rains only once tomorrow…

---

Sir Thomas Bayes (1763)
http://en.wikipedia.org/wiki/Bayesian_probability
Foundation of Probability

- **Random Variable** $A$ indicates an event that has intrinsic degree of uncertainty if $A$ occurs or not.
- $A$ can be seen as a function that chooses a value from the event space according to probability distribution $P(A)$ where an event is a subset of sample space.

- **Discrete Boolean Random Variable**
  - Sample space \{True, False\}

- **Discrete Multivalued Random Variable** (beyond PL)
  - Sample space \{Mon, ..., Sun\}

- **Continuous Random Variable** (beyond PL)
  - Sample space $\{x \in \mathbb{R}\}$

---

**Foundation of Probability**

- **The axioms of probability**
  - $P(A) \geq 0$ and $P(A) \in \mathbb{R}$ (non-negative real)
  - $P(\Omega) = 1$ (unit measure)
  - $P(A_1 \lor A_2 \ldots) = \sum_i P(A_i)$ (additivity)
  - $P(\text{true}) = 1$
  - $P(\text{false}) = 0$
  - $0 \leq P(A) \leq 1$
  - $P(A \lor B) = P(A) + P(B) - P(A \land B)$

- **Total probability theorem**
  - $P(A) + P(\neg A) = 1$ given \{true, false\}
  - $\sum_{A=v_i} P(A) = 1$ given \{\textit{v}_1, \ldots, \textit{v}_k\}
### Joint Distribution

- Probability of multiple events (A and B) in conjunction

\[ P(A \land B) = P(A,B) \]

\[ P(A \land B \land C \land D \ldots) \]

\[ P(X,Y) \]

\[ P(X=CS,Y=Yes) = P(CS, Yes)? \]

\[ P(Math, No)? \]

\[ P(History, Yes)? \]

### Marginal Distribution

- Probability of one event (A) regardless of other events (B)

- Derived from a joint distribution by integrating/summing out

\[ P(A) = P(A \land B) + P(A \land \neg B) \]

\[ P(X,Z,\ldots) = \sum_Z P(X,\theta,Z,\ldots) \]

- Sum Rule

\[ P(A) = \sum_{B=v_i} P(A \land B), \quad B = \{v_i | i = 1, \ldots, n\} \]

\[ P(Yes)? = P(CS, Yes) + P(Math, Yes) + P(History, Yes) = \frac{1}{8} \]

\[ P(CS)? = P(CS, Yes) + P(CS, No) = \frac{3}{8} \]
Conditional Distribution

- Probability of one event (A) given the other event (B)
- Derived by combining a joint distribution and a marginal distribution
  
  \[ P(A|B) = \frac{P(A \cap B)}{P(B)} \]

  \[ P(\text{Yes}|\text{CS}) = \frac{P(\text{Yes} \cap \text{CS})}{P(\text{CS})} = \frac{\frac{4}{9}}{\frac{3}{9}} = \frac{2}{3} \]

  \[ X = \text{College Major} \]
  \[ Y = \text{Likes “Games”} \]

- Product Rule
  
  \[ P(A \cap B) = P(A|B)P(B) \]
  
  \[ P(A \cap B) = P(B|A)P(A) \]

- Chain Rule (factorization)
  
  \[ P(A \cap B \cap C) = P(A|B \cap C)P(B|C)P(C) \]

Bayes Rule

- Product rule is symmetric
  
  \[ P(A \cap B) = P(A|B)P(B) \]
  
  \[ P(A \cap B) = P(B|A)P(A) \]

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

- This leads to the following form known as Bayes Rule
How to use Bayes Rule

- \textit{h}: hypothesis (A)
- \textit{d}: evidence (B)
- Compute \( P(h|d) \) given \( P(d|h) \) and \( P(h) \) known

\[
p(h | d) = \frac{p(d | h) p(h)}{\sum_{h' \in H} p(d | h') p(h')}
\]

Bayesian Inference

- I’ve got this evidence, what’s the chance that this hypothesis is true?
  - prior probability \( P(h) \): chance of \( h \) to be true without any evidence
  - a new evidence (data) ‘\( d \)’: more information about ‘\( h \)’
  - posterior probability \( P(h|d) \)
- Inference is to compute the posterior!!!
  - I’ve got a stomach pain, how likely that I have appendix?
  - Given NY Dow Jones index goes down by 5%, what is the chance of the amount of foreclosures increasing this month?
  - I see the traffic signal goes green and no cars are visible approaching from right and left, what is the chance of me getting involved in an accident by crossing the intersection?
More on Bayesian Inference

• What this Bayes rule tell us?
• You get high posterior when:
  – Hypothesis is plausible: high \( P(h) \)
  – Hypothesis strongly predicts the observed evidence/data: high \( P(d|h) \)
  – Evidence/data is very surprising: low \( P(d) \)

\[
P(h|d) = \frac{P(d|h)P(h)}{P(d)}
\]

posterior \( \propto \) likelihood \( \times \) prior

Example

• One day you wake up with a headache: You think “Noo! 50% of flues are associated with headaches so I must have 50-50 chance on coming down with flu."
• What is prior? \( P(F) = \frac{1}{40} \)
• What is likelihood? \( P(H|F) = \frac{1}{2} \)
• What is posterior? \[
P(F|H) = \frac{\frac{1}{2} \cdot \frac{1}{10}}{\frac{1}{2} \cdot \frac{1}{10} + \frac{1}{40}} = \frac{1}{5}
\]

Inference: what is the prob. of flu given a headache evidence
Statistical Modeling

• Nice formulae! but it requires you to provide numerical values for all the distributions: prior and likelihood all that...

• So you collect data and estimate these distributions from the data.
  – Interview appendix patients then ask how many had headache
  – Call randomly 1000 people and ask how many had appendix
  – Count times that DowJones went down by more than 5% given the times when foreclosure rate was coming down

• This is called Statistical Modeling

More on Statistical Modeling

• One caveat: population can be huge!!!
• You cannot get all data!!!

• Strategy
  – Data = Statistical Samples drawn from a distribution
  – Sampling process assumption:
  – IID, independent and identically distributed
  – Estimate the distribution from the samples or
  – Estimate population statistics from sample statistics
Two Approach in Statistical Modeling

• Do you have statistical insights of the underlying distribution?
  – Counting flip-coins ===== Binomial distribution
  – Weight of SF residents ===== Normal distribution

• No: Non-Parametric Modeling
  – Histogramming and more…

• Yes: Parametric Modeling (2 steps)
  – Find a parameterize function for the known distribution
  – Solve a parameter estimation problem by fitting the function to data

Joint Distribution is GREAT!!!

• Hassle to model every different distributions, right?
• You have a problem with $N$ random variables
• Statistically model a $N$-variate joint distribution $P(X_1,..,X_N)$
• You can derive the probability of any logical expression from this joint distribution !!!

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

• Use various inferential rules
  – Sum Rule: $P(A,B,C,\ldots) \rightarrow P(A,B), P(A), P(B)$
  – Product Rule: $P(A,B) \rightarrow P(A|B)P(B)$
  – Chain Rule: $P(A,B,C,\ldots,Z) \rightarrow P(A|B,C,\ldots,Z) P(B|C,\ldots,Z)\ldots P(Z)$
  – Bayes Rule: $P(A|B) \rightarrow P(B|A)$

• E.g., you can derive $P(A)$, $P(B)$, $P(A|B)$, $P(B|A)$ from $P(A,B)$
Joint Dist. is TERRIBLE!!!

- Well here is a catch…
- **Impossible to create it for more than small number of variables**
  - $M=10$ variables = $2^{10}$ entries for Boolean RV
  - $M=10$ variables = $N^{10}$ entries for $N$-valued RV (8-bit pixel)

  $M$: number of variables
  - $M$ can be a number of all stocks in the market
  - $M$ can be all possible pixel locations for a video tracker
  - $M$ can indicate # of all stars in the universe
  - $M$ can represent # of all neurons in our brain
  - $M$ can be all users in twitter

Statistical Independence

- Some domain knowledge make things better
- If $A, B$ are **statistically independent** then:
  - $P(A \land B) = P(A, B) = P(A)P(B)$
  - $P(A|B) = P(A)$
  - $P(B|A) = P(B)$
  - **A**: scores of SF Giants in the next game
  - **B**: scores of my exam and I am not a Giants fan ($A \perp B$)

- If $A, \ldots, Z$ are **mutually independent** then:
  - $P(A \land B \land \ldots \land Z) = P(A)P(B) \ldots P(Z)$
  - "Naïve Assumption"
  - $2^n$ entries --- 52 entries (for Boolean)
More on Statistical Independence

• Given three events A, B, C
• A and B can be independent once you know C is true/false

• Conditional Independence
• \[ P(A \land B|C) = P(A|C)P(B|C) \]
• \[ P(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n|\mathbf{d}) = P(\mathbf{y}_1|\mathbf{d}) \cdot P(\mathbf{y}_2|\mathbf{d}) \cdots P(\mathbf{y}_n|\mathbf{d}) \]

• Applications
  – Naïve Bayesian Classifier
  – Bayesian Network

What for?

• Providing probabilistically sound algorithms for realizing useful functions (PF)
  – Classification
  – Regression
  – Density estimation

\[ \text{Bayesian Reasoning} \quad \text{Deterministic} \quad \text{Probabilistic} \]

\[ \begin{align*}
\text{Input Attributes} & \quad \text{Classifier} & \quad \text{Prediction of categorical output} \\
\text{Input Attributes} & \quad \text{Density Estimator} & \quad \text{Probability} \\
\text{Input Attributes} & \quad \text{Regressor} & \quad \text{Prediction of real-valued output}
\end{align*} \]
PF: Bayesian Classification

- Hypothesis space $H$ consists of $K$ classes $h_k$
- Get data 'd' as an input feature
- **Which class does this input belong to???

1) Compute post. dist. $P(h_k \mid d)$ for all $h_k \in H$
2) Take the class 'k' with the highest posterior with d

- Formally

$$k^* = \arg\max_k P(h_k \in H \mid d)$$

Bayes Rule

$$k^* = \arg\max_k \frac{P(d \mid h_k)P(h_k)}{P(d)}$$ Simplify

$$P(d \mid h_j)P(h_j)$$

PF: Bayesian Classification
**PF: Naïve Bayesian Classifier**

- Typically we work with multiple features
  \[ k^* = \arg\max_k P(h_k | d_1, \ldots, d_M) \]
  \[ k^* = \arg\max_k \frac{P(d_1, \ldots, d_M | h_k)P(h_k)}{P(d_1, \ldots, d_M)} \]
  \[ k^* = \arg\max_k \prod_{m=1}^{M} P(d_m | h_k)P(h_k) \]
- But cannot easily build the joint dist so…
- This is much more **practical** (Data Mining)!

**PF: Bayesian Regression**

- A function \( f: X \mapsto Y \)
- A parameterized model \( y = f(x, w) = wx + \text{noise} \)
- Noise is independent, 0 mean, variance \( \sigma^2 \)
- \( P(y|w, x) \) models this function probabilistically
- Get a set of IID data samples \( \{x_m, y_m\}, m = 1, \ldots, M \)
- **Find parameters `w` that make the model fit to the data most**

<table>
<thead>
<tr>
<th>input</th>
<th>output</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2.2</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>1.5</td>
<td>1.9</td>
</tr>
<tr>
<td>4</td>
<td>3.1</td>
</tr>
</tbody>
</table>
Bayesian Regression

- Bayesian Regression
  \[ w^* = \arg \max_w P(w | (x_1, y_1), \ldots, (x_M, y_M)) \]

- Maximum Likelihood Regression
  \[ w^* = \arg \max_w P((y_1, \ldots, y_M) | w, (x_1, \ldots, x_M)) \]
  \[ w^* = \arg \max_w \prod_m P(y_m | w, x_m) \quad \text{IID} \]
  \[ w^* = \arg \max_w \prod_m \exp \left( -\frac{1}{2} \frac{(y_m - wx_m)^2}{\sigma} \right) \quad \text{Normal} \]
  \[ w^* = \arg \max_w \sum_m -\frac{1}{2} \frac{(y_m - wx_m)^2}{\sigma} \quad \text{Log e} \]
  \[ w^* = \arg \min_w \sum_m (y_m - wx_m)^2 \quad \text{Simplify} \]

So the maximum likelihood regression
\[ w^* = \arg \max_w P((y_1, \ldots, y_M) | w, (x_1, \ldots, x_M)) \]
Is the same as this...
\[ w^* = \arg \min_w \sum_m (y_m - wx_m)^2 = \sum_m \text{error}_m^2 \]
This is known as least squares problem

Least Squares (Linear) Regression is Maximum Likelihood Regression when we model the PDF function by Gaussian/Normal distribution !!!
Bayesian Learning

• Bayesian learning is an estimation of probability distribution given a data set (as evidence)

• E.g., counting frequency from

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Math</td>
<td>Yes</td>
</tr>
<tr>
<td>CS</td>
<td>No</td>
</tr>
<tr>
<td>Math</td>
<td>No</td>
</tr>
<tr>
<td>CS</td>
<td>Yes</td>
</tr>
<tr>
<td>History</td>
<td>No</td>
</tr>
<tr>
<td>Math</td>
<td>Yes</td>
</tr>
</tbody>
</table>

• E.g., making histogram

• E.g., estimating a function by least squares

\[
P(d|h_1)P(h_1) + P(d|h_2)P(h_2) + P(d|h_3)P(h_3)\]

Any Other Reasoning with Uncertainty?

• Oh Yes!
  – Fuzzy Logic
  – Dempster-Shafer
  – Three-valued Logic
  – Non-monotonic Logic

• But “probability theory” is the one of the methods that is most sound mathematically
  – If you are gambling using them, you can’t be unfairly exploited by an opponent using some other system [di Finetti 1931]
Summary

• Probability Theory for Reasoning with Uncertainty
  – Bayesian Inference
  – Statistical Modeling
  – Bayesian Classification
  – Bayesian Regression
  – Bayesian Learning

• Next
  – Non-Parametric Statistical Modeling
  – Kernel Density Estimation
  – Mean Shift
  – FP#1 on Eigenface (2nd Week)