Note

- 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 5 papers** on your subtopic.
  - **Submit your subtopic/summary/selected papers to Canvas forum** for my review (~~make your own thread~~).
  - **Due on March 18**

PF

Framework for Bayesian Probabilistic Reasoning

CSC 872

Pattern Analysis and Machine Intelligence

Logical Reasoning: Review

Knowledge based agent

- 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** Probability
  - 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)
An essay toward solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society of London, 53: 370-418
http://en.wikipedia.org/wiki/Bayesian_probability

Foundation of Probability

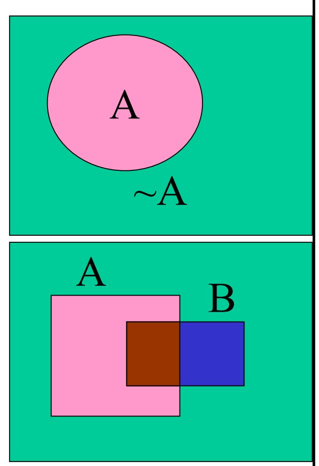
X: {true, false}
P(X)

- **Random Variable $\mathbf{A}$** indicates an event that has intrinsic degree of **uncertainty** if $\mathbf{A}$ occurs or not
- $\mathbf{A}$ can be seen as a **function** that chooses a value from the **event space** according to probability distribution $\mathbf{P(A)}$ 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 $\{x \in \mathbf{R}\}$

Foundation of Probability

- **The axioms of probability**
  - $\mathbf{P(A) \geq 0}$ **and** $\mathbf{P(A) \in R}$ **(non-negative real)**
  - $\mathbf{P(\Omega) = 1}$ **(unit measure)**
  - $\mathbf{P(A_1 \vee A_2 \vee \ldots) = \sum_i P(A_i)}$ **(additivity)**

  - $\mathbf{P(True) = 1}$
  - $\mathbf{P(False) = 0}$
  - $\mathbf{0 \leq P(A) \leq 1}$
  - $\mathbf{P(A \vee B) = P(A) + P(B) - P(A \wedge B)}$

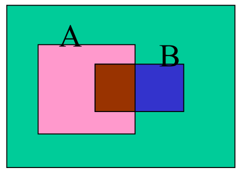


- **Total probability theorem**
  - $\mathbf{P(A) + P(\sim A) = 1}$ **given {true,false}**
  - $\mathbf{\sum_{A=v_i} P(A) = 1}$ **given** $\mathbf{\{v_1,\ldots,v_K\}}$

Joint Distribution

Probability

- Probability of multiple events (A and B) in conjunction
  - *P(A ^ B) := P(A, B)*



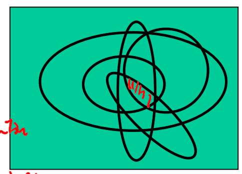
X = {CS, Math, History}
Y = {Yes, No}

X = College Major

Y = Likes "Games"

X	Y
Math	Yes
CS	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

  - *P(A ^ B ^ C ^ D,…)*



distribution

P(X,Y)

probability value

P(~~X~~=CS,~~Y~~=Yes) = P(CS, Yes)? 1/4
P(Math, No)? 1/4
P(History, Yes)? 0

Marginal Distribution

- Probability of one event (A) regardless of other events (B)
- Derived from a joint distribution by integrating/summing out
  - $P(A) = P(A \wedge \overset{=\text{True}}{B}) + P(A \wedge \overset{=\text{False}}{\sim B})$

X = College Major

Y = Likes "Games"

X	Y
Math	Yes
CS	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

$P(x, y, z, \ldots)$

- **Sum Rule** $P(x, z, \ldots) = \sum_y P(x, y, z, \ldots)$
  - $P(A) = \sum_{B=v_i} P(A \wedge B)$ $= P(A, B=v_1) + P(A, B=v_2) + \cdots + P(A, B=v_n)$
  - P(Yes)? $= \underbrace{P(CS, Yes)}_{1/4} + \underbrace{P(Math, Yes)}_{1/4} + \underbrace{P(Hist, Yes)}_{0} = 1/2$
  - P(CS)? $= \underbrace{P(CS, Yes)}_{2/8} + \underbrace{P(CS, No)}_{1/8} = \frac{3}{8}$

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) = P(A ^ B) / P(B)*
  - P(Yes|CS) ? $= \frac{P(CS, Yes)}{P(CS)} = \frac{1/4}{3/8} = \frac{2}{3}$

X = College Major

Y = Likes "Games"

X	Y
Math	Yes
CS	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

- **Product Rule**
  - *P(A ^ B) = P(A|B) P(B)*

    $P(B \wedge A) = P(B|A) P(A)$
- **Chain Rule (factorization)**
  - *P(A ^ B ^ C) = P(A|B ^ C) P(B|C) P(C)*

    $P(A \wedge B \wedge C \wedge D) = P(A|B \wedge C \wedge D) P(B|C \wedge D) P(C|D) P(D)$

Bayes Rule

- **Product rule is symmetric** $P(A, B) = P(B, A)$
  - *P(A ^ B) = P(A|B) P(B)*
  - *P(A ^ B) = P(B|A) P(A)*
- **This leads to the following form known as Bayes Rule**
  - ***P(A|B) = P(B|A) P(A) / P(B)***

How to use Bayes Rule

- *h*: hypothesis (A) — Query / Question — α
- *d*: evidence (B) — data — K.B.
- Compute *P(h|d)* given *P(d|h) and P(h)* known

KB → α

Posterior probability — α

Likelihood

Prior probability — KB

$$p(h \mid d) = \frac{p(d \mid h)\, p(h)}{\sum_{h' \in H} p(d \mid h')\, p(h')} = p(d)$$

Decision Making!

Sum over space of hypotheses

Bayesian Inference

$P(h) \xrightarrow[\text{evidence, Perception}]{d} P(h|d)$

best we can do without "d"

- **I've got this evidence, what's the chance that this hypothesis is true?**
  - a hypothesis '***h***'
  - **prior** probability ***P(h)*****:** chance of *h* being 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 appendicitis?
  - 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)**

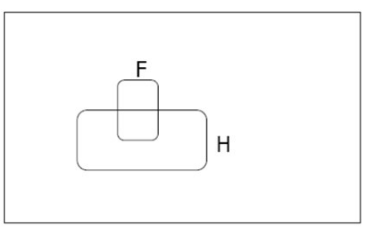
rare

likelihood prior KB,

$$P(h|d) \propto = \frac{P(d|h)P(h)}{P(d)}$$

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

Example



H = "Have a headache"
F = "Coming down with Flu"

P(H) = 1/10
P(F) = 1/40
P(H|F) = 1/2

- One day you wake up with a headache: You think "Nooo! 50% of flues are associated with headaches so I must have 50-50 chance on coming down with flu..
- What is prior? $P(F) = 1/40$
- What is likelihood? $P(H|F) = 1/2$
- What is posterior? $P(F|H) = \frac{1/2 \cdot 1/40}{1/10} = \frac{1}{8}$
- Inference: what is the prob. of flu given a headache evidence

Statistical Modeling

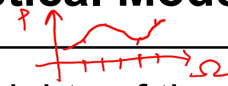
- Nice formulae ! but it **requires** you to provide numerical values for all the distributions: prior and likelihood all that…
- So you **collect data** and **estimate these distributions from the data**.
  - Interview appendicitis patients then ask how many had headache
  - Call randomly 1000 people and ask how many had appendicitis
  - Count times that DowJones went down by more than 5% given the times when foreclosure rate was coming down
- This is called "**Statistical Modeling**"

K, B,

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? Shape ?
  - 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 !!!**

$P(h|d) = \frac{P(d|h)P(h)}{P(d)}$

- Use various inferential rules
  - Sum Rule: $P(A,\cancel{B},\cancel{C},\ldots)$ → $P(A,B)$, $P(A)$, $P(B)$
  - Product Rule: $P(A,B)$ → $P(A|B)P(B)$
  - Chain Rule: $P(A,B,C,..,Z)$ → $P(A|B,C,..,Z)\ P(B|C,..,Z)\ldots P(Z)$
  - Bayes Rule: $P(A|B)$ → $P(B|A)$
- 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… Curse of dimensionality
- ***Impossible to create it for more than small number of variables*** 2^n $P(x_1, x_2, x_3, \dots, x_n)$
  - M =10 variables = $\mathbf{2^{10}}$ 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 Tiktok

Statistical Independence

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

- If *A,..,Z* are mutually independent then:
  - $P(A \wedge B \wedge \dots \wedge Z) = P(A)P(B)\dots P(Z)$ $2^n \hookrightarrow 2n$
  - **"Naïve Assumption"**
  - **2^{26} 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 ^ B|C) = P(A|C)P(B|C)***

$$P(x_1, x_2 \cdots x_n | d) = P(x_1|d) \cdot P(x_2|d) \cdots P(x_n|d)$$

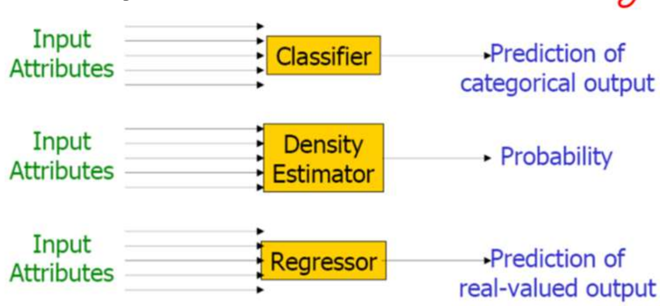
$$= \prod_{i=1}^{n} P(x_i|d)$$

- **Applications**
  – ***Naïve Bayesian Classifier***
  – ***Bayesian Network***

What for?

Bayesian Reasoning
Posterior Prob $\propto$ Prior Prob

- Providing probabilistically sound algorithms for realizing useful functions (PF)
  – **Classification**
  – **Regression**
  – **Density estimation** = Statistical Modeling



PF: Bayesian Classification

$H = \{h_1, \dots h_k, \dots h_K\}$

- Hypothesis space ***H*** consists of ***N*** classes $\boldsymbol{h_i}$
- Get data '***d***' as an input feature
- **Which class does this input belong to???**

$P(h_1|d), P(h_2|d), \dots P(h_K|d)$

1) Compute posterior P($\boldsymbol{h_i}$ | ***d***) for all $\boldsymbol{h_i} \in \boldsymbol{H}$
2) Take the class '***i***' with the highest posterior of ***d***

- Formally

$$k^* = \text{argmax}_i P(h_k \in H|d)$$

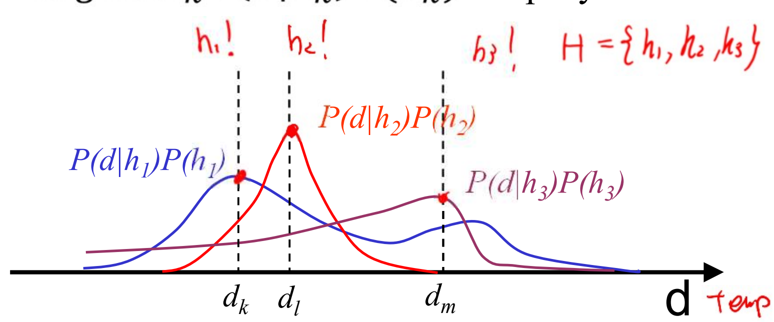
result

PF: Bayesian Classification

$$k^* = argmax_k P(h_k|d)$$

$$k^* = argmax_k \frac{P(d|h_k)P(h_k)}{P(d)} \text{ Bayes Rule}$$

$$k^* = argmax_k P(d|h_k)P(h_k) \text{ Simplify}$$



PF: Naïve Bayesian Classifier

inputs

- Typically we work with multiple features

$$k^* = argmax_k P(h_k|d_1,..,d_M)$$

$$k^* = argmax_k \frac{P(d_1,..,d_M|h_k)P(h_k)}{P(\cancel{d_1,..,d_M})}$$

$$k^* = argmax_k P(d_1,..,d_M|h_k)P(h_k)$$

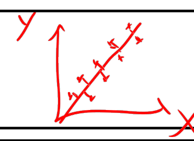
$P(d_1|h_i)\,P(d_2|h_i)\cdots P(d_n|h_i)\,P(h_i)$

- But cannot easily build the joint dist so…

$$k^* = argmax_k \prod_m P(d_m|h_k)P(h_k)$$

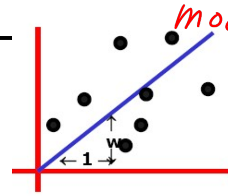
- This is much more **practical** (Data Mining) !

PF: Bayesian Regression



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

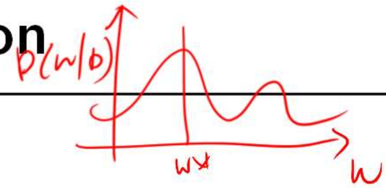
model fitting



DATASET

inputs	outputs
$x_1 = 1$	$y_1 = 1$
$x_2 = 3$	$y_2 = 2.2$
$x_3 = 2$	$y_3 = 2$
$x_4 = 1.5$	$y_4 = 1.9$
$x_5 = 4$	$y_5 = 3.1$

Bayesian Regression



- Bayesian Regression

$$w^* = argmax_w P(w|(x_1, y_1), .., (x_M, y_M))$$

- **Maximum Likelihood Regression**

$$w^* = argmax_w P((y_1, .., y_M)|w, (x_1, .., x_M))$$

$$w^* = argmax_w \prod_m P(y_m|w, x_m) \qquad \text{IID} \quad \text{naive}$$

$$w^* = argmax_w \prod_m \exp(-\tfrac{1}{2}(\tfrac{y_m - wx_m}{\sigma})^2) \qquad \text{Normal}$$

$\underset{x}{argmax} f(x) = \underset{x}{argmax}\, g(f(x))$ when g is monotonic

$$w^* = argmax_w \sum_m -\tfrac{1}{2}(\tfrac{y_m - wx_m}{\sigma})^2 \qquad \text{Log}$$

$$\boxed{w^* = argmin_w \sum_m (y_m - wx_m)^2 \qquad \text{Simplify}}$$

Bayesian Regression

- So the maximum likelihood regression

  $w^* = argmax_w P((y_1, .., y_M)|w, (x_1, .., x_M))$
- Is the same as this…

  $w^* = argmin_w \sum_m (y_m - wx_m)^2 = \sum_m error_m^2$
- 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 !!!**

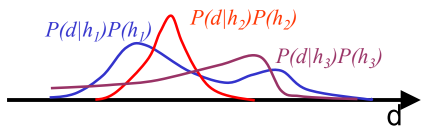
Bayesian Learning

Bayesian KB.

- Bayesian learning is an estimation of probability distribution given a data set (as evidence)
- E.g., counting frequency from

X	Y
Math	Yes
CS	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

- E.g., making histogram



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

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]
  - http://en.wikipedia.org/wiki/Bruno_de_Finetti

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 Session)**