Bayesian Theory and Bayesian Modeling

Hsing-Kuo Pao (鮑興國)
National Taiwan University of Science & Technology (Taiwan Tech)

Outline

- Introduction and Fundamentals
- Frequentist vs. Bayesian
- Bayes Rule
- Maximum Likelihood Estimation, Maximum A Posteriori Estimation
- Bayesian Prediction

Introduction and Fundamentals

— Bayesian Theory and Bayesian Modeling

What is Learning?

- **Definition**: A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience.

- To work on Induction?
  - Data + **Inductive bias** $\Rightarrow$ model (a deductive logic)
  - E.g.: Data points + Linear assumption $\Rightarrow$ linear model!

Aristotelian logic

- If A is true, then B is true
- A is true
- Therefore, B is true

E.g. 1

- A: The class is canceled
- B: Professor does not show up

Real-world is uncertain

E.g. 2

- A: It is raining
- B: The grass is wet

Problems with pure logic:

- Don't have perfect information (missing attributes)
- Don't really know the model (not sure the model type)
- Model is non-deterministic

Why not build a logic of uncertainty!

Probabilistic Approach for Uncertainty Modeling

- Probabilistic prediction
  - calculate explicit probabilities for hypothesis.
  - e.g.: this pneumonia patient has a 93% chance of complete recovery.
- Predict multiple hypotheses, weighted by their (posterior) probabilities
- Resistance to noisy data: in Bayesian modeling, each example can increase/decrease probability that certain hypothesis h is correct, instead of ruling out any inconsistent hypotheses

Adding Prior Knowledge

- Final Model: Prior knowledge + Observed data
  - Combining prior and data: final prob. of h
- In Bayesian learning, everybody may have different "opinions"
- Before now, probability may simply mean frequency!

"Probability theory is nothing more than common sense reduced to calculation."

- Pierre-Simon Laplace, 1814

"Probability does not exist."

- De Finetti

General Difficulties of Bayesian Learning

- Require large initial knowledge of many probability
  - Often estimated in practice
- Large computational costs
  - Linear to # of hypotheses
  - Can be reduced in certain situations
- Even when intractable
  - Give a standard of optimal decision making against which other methods can be measured

Frequentist vs. Bayesian

— Bayesian Theory and Bayesian Modeling

Frequentist vs. Bayesian

- Frequentist statistics
  - a.k.a. "orthodox statistics"
  - Probability = frequency of occurrences in infinite # of trials
  - Arose from sciences with populations
  - *p*-values, *t*-tests, ANOVA, etc.
- Frequentist vs. Bayesian debates have been long and acrimonious

Frequentist vs. Bayesian (cont'd)

- "In academia, the Bayesian revolution is on the verge of becoming the majority viewpoint, which would have been unthinkable 10 years ago."
  - Bradley P. Carlin, professor of public health, University of Minnesota
    (New York Times, Jan 20, 2004)

Frequentist vs. Bayesian (cont'd)

- If necessary, please leave these assumptions behind (for this lecture):
  - "A probability is a frequency"
  - "Probability theory only applies to large populations"
  - "Probability theory is arcane and boring"

Bayes Rule

— Bayesian Theory and Bayesian Modeling

Tossing a Coin?

- Consider the probability of whether a coin toss will land on heads (or tails)
- Problem 1: After 100 tosses, we find out that 51 times the coin land with head up.

  What is your estimation of $P(\theta) = P(\text{head})$?

➲ Answer = 51/100?

- Problem 2: After 2 tosses, we find out that 2 times the coin land with head up.

  What is your estimation of $P(\theta) = P(\text{head})$?

➲ Answer = 2/2?

- Suppose that you know more information…

Bayes Rule

- **Thomas Bayes (1763)** “An essay towards solving a problem in the doctrine of chances”. *Philosophical Transactions of the Royal Society of London,* **53**: pp. 370-418.
- Best h in $\mathcal{H}$ given $\mathcal{D}$
  - best = most probable
  - BR: direct method of computing it

$$P(h \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid h)P(h)}{P(\mathcal{D})}$$

- Notation

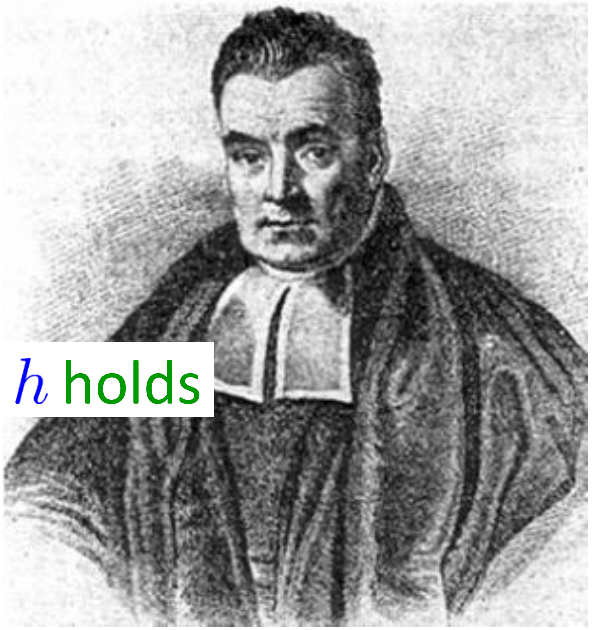
  prior probability

  - $P(h)$: prob. h (hypothesis) holds
  - $P(\mathcal{D})$: prob. $\mathcal{D}$ (dat[illegible]erved

  likehood

  - $\mathcal{L}(\mathcal{D} \mid h) = P(\mathcal{D} \mid h)$: $\mathcal{D}$ observed when h holds
  - $P(h \mid \mathcal{D})$: h holds when $\mathcal{D}$ observed (!)

  posterior probability



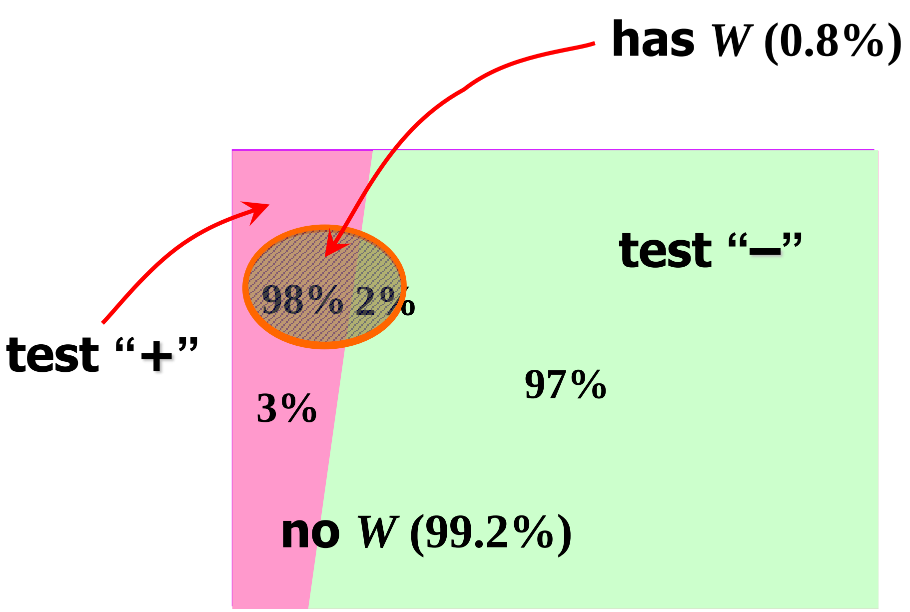
Usage of Bayes Rule

- Bayes rule: $$P(h \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid h)P(h)}{P(\mathcal{D})}$$

  the realm of density estimation
- Generally, best h maximizes $P(h \mid \mathcal{D})$
  - MAP: *Maximum A Posteriori*
  - $h_{\mathrm{MAP}} = \mathrm{argmax}_h P(h \mid \mathcal{D})$
- Especially, if every h equally likely
  - ML: *Maximum Likelihood*
  - $h_{\mathrm{ML}} = \mathrm{argmax}_h P(\mathcal{D} \mid h)$
- Note: applicable to general h & $\mathcal{D}$

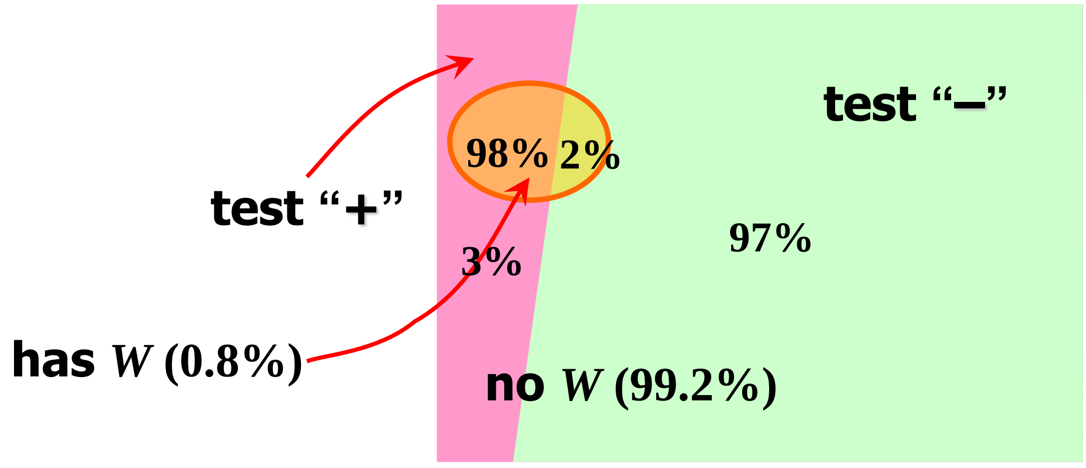
Example

- $f(x) =$ lab test for disease W
  - Return "+" in 98% of cases where x really has
  - Return "–" in 97% of cases where really not
- Prior knowledge
  - 0.8% of population has W
- $f(x) = +$
  - what to believe?



Maximum Likelihood Estimation

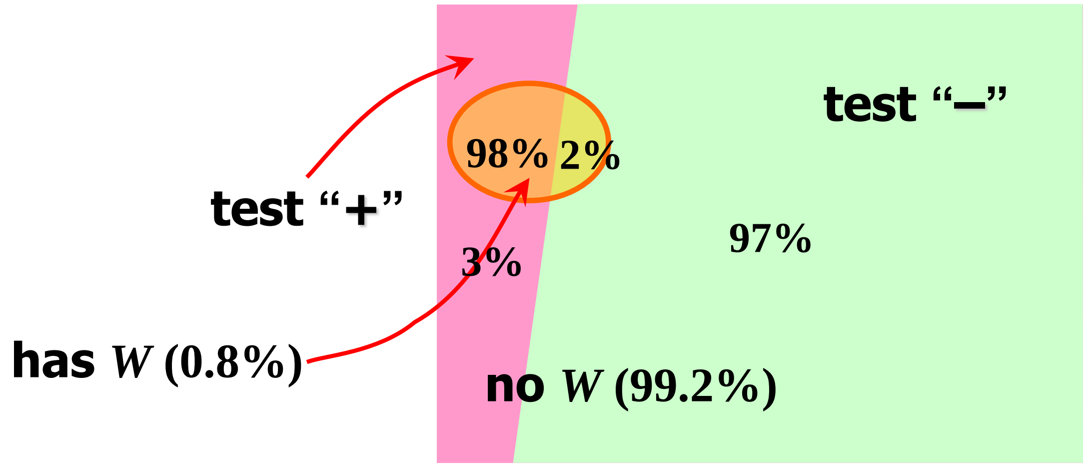
- $$\begin{aligned} h_{\mathrm{ML}} &= \mathrm{argmax}_h P(D \mid h) \\ &= \mathrm{argmax}_h \{P(+ \mid cancer), P(+ \mid \neg cancer)\} \\ &= \mathrm{argmax}_h \{0.98, 0.03\} \\ &= cancer \end{aligned}$$



Maximum A Posteriori Estimation

- $$\begin{aligned} P(cancer \mid +) &= \frac{P(+ \mid cancer)P(cancer)}{P(+)} \\ &= \frac{0.98 \cdot 0.008}{P(+)} = 0.0078/P(+) \\ P(\neg cancer \mid +) &= \frac{P(+ \mid \neg cancer)P(\neg cancer)}{P(+)} \\ &= \frac{0.03 \cdot 0.992}{P(+)} = 0.0298/P(+) \end{aligned}$$

- $h_{\text{MAP}} = \neg cancer$



Notes from Example

- $P(+)$ can be computed
  - not known in advance
  - an indirect way:
    probabilities $P(cancer \mid +), P(\neg cancer \mid +)$ sum up to 1
- Posterior probability of $cancer \gg$ *a priori*
  - still, the most probable hypothesis is that the patient does not have cancer
- Result depends strongly on *a priori* probabilities
  - must be known
- Hypotheses are not 100% accepted or rejected!

Coin Tossing

- Let $X = \{0, 1\}$ be the binary variable that represents the result of a coin toss, $X = 1$ if coin land with heads up and $X = 0$ otherwise. Let $\theta = P(X = 1)$. Then, we say that X a random variable that follows Bernoulli distribution with parameter θ

$$X \sim \mathrm{Be}(\theta)$$

- The probability of observe a coin toss outcome x (either head of tail) given θ is

$$P(x|\theta) = \theta^x(1-\theta)^{1-x}$$

Maximum Likelihood Estimation – Continuous Hypotheses

- Suppose we observed a sample of size n from the above distribution $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$. Assuming that samples are independently and identically distributed (iid), the joint probability of $\mathbf{x}$ as a function of parameter θ is called likelihood $\mathcal{L}(\mathbf{x} \mid \theta)$.

  We want to find a setting of θ that maximizes the likelihood function.

$$\hat{\theta} = \mathrm{argmax}_{\theta} \mathcal{L}(\mathbf{x}|\theta)$$

1. Write down the likelihood function

$$\mathcal{L}(\mathbf{x}|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}$$

2. Take $\log$ of the likelihood function

$$\begin{aligned} \ell(\theta) &= \sum_{i=1}^{n} \log\left[\theta^{x_i}(1-\theta)^{1-x_i}\right] \\ &= \left(\sum_{i=1}^{n} x_i\right)\log\theta + \left(n - \sum_{i=1}^{n} x_i\right)\log(1-\theta) \end{aligned}$$

Maximum Likelihood Estimation (cont'd)

3. Take the derivative of log likelihood w.r.t. θ and set to zero

$$\frac{\partial l(\mathbf{x}|\theta)}{\partial \theta} = \left(\sum_{i=1}^{n} x_i\right) \frac{\partial \log(\theta)}{\partial \theta} + \left(n - \sum_{i=1}^{n} x_i\right) \frac{\partial \log(1-\theta)}{\partial \theta} = 0$$

4. Solve the equation, we have

$$\hat{\theta} = \frac{\sum_{i=1}^{n} x_i}{n}$$

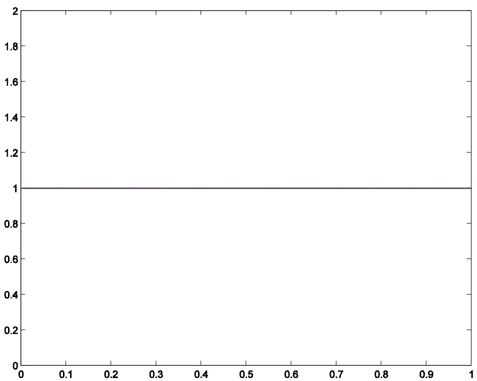
Applying Bayes Rule Iteratively

$$p(h) = 1 \,|_0^1$$

$$p(h|x_1) = \frac{p(x_1|h)p(h)}{p(x_1)} = \frac{\theta \cdot 1}{\int_\theta \theta d\theta} = 2\theta \,|_0^1$$

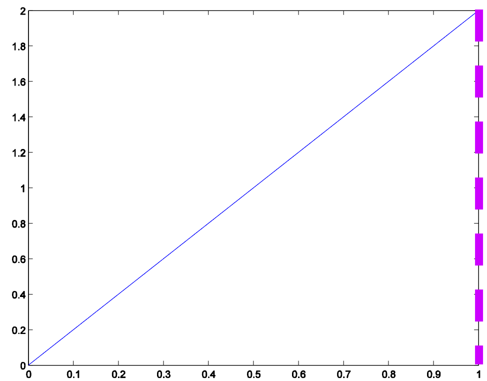
$$p(h|x_1, x_2) = \frac{p(x_2|h, x_1)p(h|x_1)}{p(x_2|x_1)} = \frac{(1-\theta)\theta}{p(x_2, x_1)/p(x_1)} = 6\theta(1-\theta) \,|_0^1$$

$p(h)$



hypothesis: θ

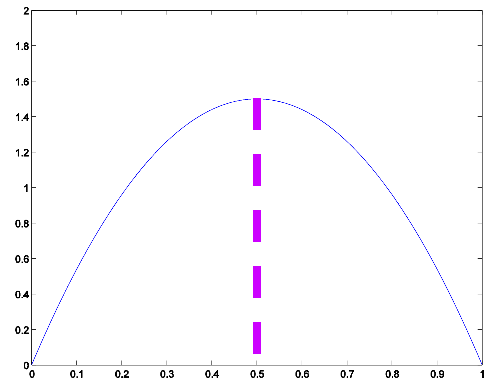
$p(h \mid x_1 = 1)$



hypothesis: θ

MAP

$p(h \mid x_1 = 1, x_2 = 0)$



hypothesis: θ

MAP

MLE on Multinomial Dist.

- The outcome of a random event is one of K mutually exclusive and exhaustive states, each with probability of P_i where $\sum_{i=1}^{K} P_i = 1$
- N trials where outcome i occurred N_i times and
  $\sum_{i=1}^{K} N_i = N$
- $P(N_1, N_2, \ldots, N_k) = N! \prod_{i=1}^{K} \frac{P_i^{N_i}}{N_i!}$
- The MLE of $\widehat{P}_i$ is $\widehat{P}_i = \dfrac{N_i}{N}$


AU6561842D0
DEUTSCHE BUNDESBANK
10
ZEHN DEUTSCHE MARK
10
AU6561842D0
DEUTSCHE
0,3
0,2
0,1
$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
0
1
2
3
4
5
6
$\mu-\sigma$
μ
$\mu+\sigma$

MLE on Gaussian Dist.

- $X = \{x^t\}_{t=1}^N$ with $x^t \sim \mathcal{N}(\mu, \sigma^2)$
- $p(x^t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x^t-\mu)^2}{2\sigma^2}\right], \quad -\infty < x^t < \infty$
- $\mathcal{L}(\mu, \sigma \mid X) = -\frac{N}{2}\log(2\pi) - N\log\sigma - \frac{\sum_t (x^t-\mu)^2}{2\sigma^2}$
- $m = \frac{\sum_t x^t}{N}, \quad s^2 = \frac{\sum_t (x^t - m)^2}{N}$
- Various estimators: MAP, Maximum likelihood, Unbiased estimators

How to Apply MLE or MAP

- By enumeration
  - e.g.: the disease testing case
  - can be computationally intractable
- By parametric methods, with a little calculus (maximization), …
  - e.g.: the coin tossing case
  - can lead to inappropriate choice of hypothesis space

Frequentist vs. Bayesian (Revisit)

- Frequentist approach:
  - by MLE
  - no assumption is necessary for building a prior, i.e., everybody gets the same answer!
  - may fail for small number of samples
- Bayesian approach:
  - by MAPE (basically)
  - need to “guess” a prior, i.e., each person may get his/her own answer!
  - can deal with small number of samples

Bayesian Classifiers and Bayesian Decision Theory

— Bayesian Theory and Bayesian Modeling

Naïve Bayes Classifier

- Text mining: given the frequency of keywords $x_1, x_2, \ldots, x_n$ in a document, predict the document class C
- Problem setting
  - Examples: attribute tuples & finite # of classes
  - $h_{\text{MAP}} = \text{argmax}_k \, P(C_k \mid x_1, \ldots, x_n)$
  - Apply Bayes theorem
- Estimates
  - (Prior) $P(C_k)$: frequencies in $\mathcal{D}$ (a dataset or a document database)
  - (Likelihood) $P(x_1, \ldots, x_n \mid C_k)$: can be VERY small for limited samples $\Rightarrow$ overfitting!

Naïve Bayes Classifier (cont'd)

- Make assumption
  - $x_1, x_2, \ldots, x_n$ are independent given C_k
  - $P(x_1, \ldots, x_n \mid C_k) = \prod_j P(x_j \mid C_k)$
- C_{NB}
  - $\text{argmax}_k \, P(C_k) \prod_j P(x_j \mid C_k)$
  - Considerably smaller amount of priors

Naïve Bayes Learning

- Compute estimates
  - $P(C_k)$: frequencies in $\mathcal{D}$
  - $P(x_j \mid C_k)$: similarly
- Compute $\operatorname{argmax}_k C_k$ for new $\mathbf{x}'$
  - If conditional independence holds, same as MAP classification
- No searching, just computation
  - Different from many iterative algorithms (e.g., neural network learning) which are common in machine learning!
  - Hypothesis space: $P(C_k), P(x_j \mid C_k)$

Some subtleties

- Independence assumption
  - Is usually violated
  - But method works anyway
  - $\arg\max$ does not require $P(C_k \mid \mathbf{x})$ is correct
- Missing or rare attribute values
  - Add "virtual" examples
  - m-estimate of probability: $\dfrac{n_c + mp}{n + m}$

    n: total no. of training examples

    n_c: no. of examples with given attr. value

    p: prior probability

    m: equivalent sample size

Recall Bayes Rule

$$P(h \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid h)P(h)}{P(\mathcal{D})}$$

- E.g. 1: (one datum) $\mathcal{D}$ is test result, h is the hypothesis "having disease w"
- E.g. 2: (a set of data) $\mathcal{D}$ is the result of n times coin tossing trials, h is the hypothesis "coin with probability θ to have head facing up"
- E.g. 3: (model building) $\mathcal{D}$ is a data set, h is simply the hypothesis/model
- E.g. 4: (several attr's) $\mathcal{D}$ is the frequency of n keywords in an article, h is the hypothesis "this article discusses about sports"

MAP Estimation Once Again

- In MAP estimation, we want to maximize

$$P(h \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid h)P(h)}{P(\mathcal{D})}$$

- Most of the time, the output h is a labeling/classification for *a new instance* (prediction) instead of a hypothesis/model/classifier for *a set of training examples* (model searching)
- Given a set of examples $\mathcal{D}$ and the most probable hypothesis $h \in \mathcal{H}$ is obtained, we can do even better for the classification on a *single* new instance

Bayes Optimal Classifier

- This far: most probable $h \in \mathcal{H}$ given $\mathcal{D}$
- More interesting: most probable *classification* for a new instance $\mathbf{x}$
  - maximize probability $P(C_k \mid \mathcal{D}, \mathbf{x})$
  - not necessarily $h_{\mathrm{MAP}}(\mathbf{x})$!
- $P(C_k \mid \mathcal{D}, \mathbf{x})$

  $$= \sum_{h_\ell} P(C_k \mid h_\ell(\mathbf{x}) P(h_\ell \mid \mathcal{D})$$
  - Bayes optimal classification takes C_k maximizing this quantity
  - A (posterior probability) weighted average of classification from all possible hypotheses

Bayes Optimal Classifier (cont'd)

- Any system
  - computing $\mathrm{argmax}_{C_k} P(C_k \mid \mathcal{D}, \mathbf{x})$ is called Bayes optimal classifier
- No other system
  - with same hypothesis space $\mathcal{H}$ & prior knowledge is better on average
- Maximizes the probability new $\mathbf{x}$ is classified correctly (given $\mathbf{x}, \mathcal{H}, \ldots$)
- "Learned" h may not be in $\mathcal{H}$!
  - $\mathcal{H}'$: comparisons on linear combinations of predictions made by $\mathcal{H}$

Gibbs Algorithm

- Optimal method needs too much prior knowledge in practice
  - $P(h_\ell \mid \mathcal{D})$: linear to $|\mathcal{H}|$
  - $P(C_k \mid h_\ell)$: linear to $|V| \cdot |\mathcal{H}|$

    V: all possible labeling, $\mathcal{H}$: all hypotheses
- Gibbs: pick random h & apply it
  - according to (estimate of) $P(h \mid \mathcal{D})$
  - surprisingly good!

Bayesian Decision Theory: Losses and Risks

- Problem: Distinguish "high-risk customer" from "good customer", medical diagnosis, earthquake prediction, etc.
- Observable evidences: $\mathbf{x} = [(x_1, x_2, \ldots, x_n]^T$

1. Choose $C = 1$ if $P(C = 1 \mid \mathbf{x}) > 0.5$

   Choose $C = 0$ otherwise

$$\Rightarrow R = 1 - \max(P(C = 1 \mid \mathbf{x}), P(C = 0 \mid \mathbf{x}))$$

2. K labels C_k, action α_ℓ, loss $\lambda_{\ell k}$ (loss incurred for taking α_ℓ when the input actually belongs to C_k

$$\begin{aligned} R(\alpha_\ell \mid \mathbf{x}) &= \sum_{k=1}^{K} \lambda_{\ell k} P(C_k \mid \mathbf{x}) \\ &= \sum_{k \neq \ell} P(C_k \mid \mathbf{x}) = 1 - P(C_\ell \mid \mathbf{x}) \end{aligned}$$

$\min(R)$:
choose the most probable $P(C_k \mid \mathbf{x})$

if $\lambda_{\ell k} = 1(\ell \neq k)$ (zero-one loss)

Loss on rejection

- For an additional action of reject (doubt)

$$\lambda_{\ell k} = \begin{cases} 0 & \text{if } \ell = k \\ \lambda & \text{if } \ell = K+1 \\ 1 & \text{otherwise} \end{cases}$$

1. $R(\alpha_{K+1} \mid \mathbf{x}) = \sum_{k=1}^{K} \lambda P(C_k \mid \mathbf{x}) = \lambda$

2. $R(\alpha_\ell \mid \mathbf{x}) = \sum_{k \neq \ell} P(C_k \mid \mathbf{x}) = 1 - P(C_i \mid \mathbf{x})$

$\rightarrow$ choose C_ℓ if $R(\alpha_\ell \mid \mathbf{x}) < R(\alpha_k \mid \mathbf{x})\ \forall k \neq \ell$ and
$R(\alpha_\ell \mid \mathbf{x}) < R(\alpha_{K+1} \mid \mathbf{x})$
or if $P(C_\ell \mid \mathbf{x}) > P(C_k \mid \mathbf{x})\ \forall k \neq \ell$ and
$P(C_\ell \mid \mathbf{x}) > 1 - \lambda$

$\rightarrow$ reject if $R(\alpha_{K+1} \mid \mathbf{x}) \leq R(\alpha_\ell \mid \mathbf{x}), \ell = 1, \ldots, K$
or if $P(C_\ell \mid \mathbf{x}) \leq 1 - \lambda, \ell = 1, \ldots, K$

Discriminant Functions

- How do we represent pattern classifiers?

$\Rightarrow$ The most common way is through discriminant functions.

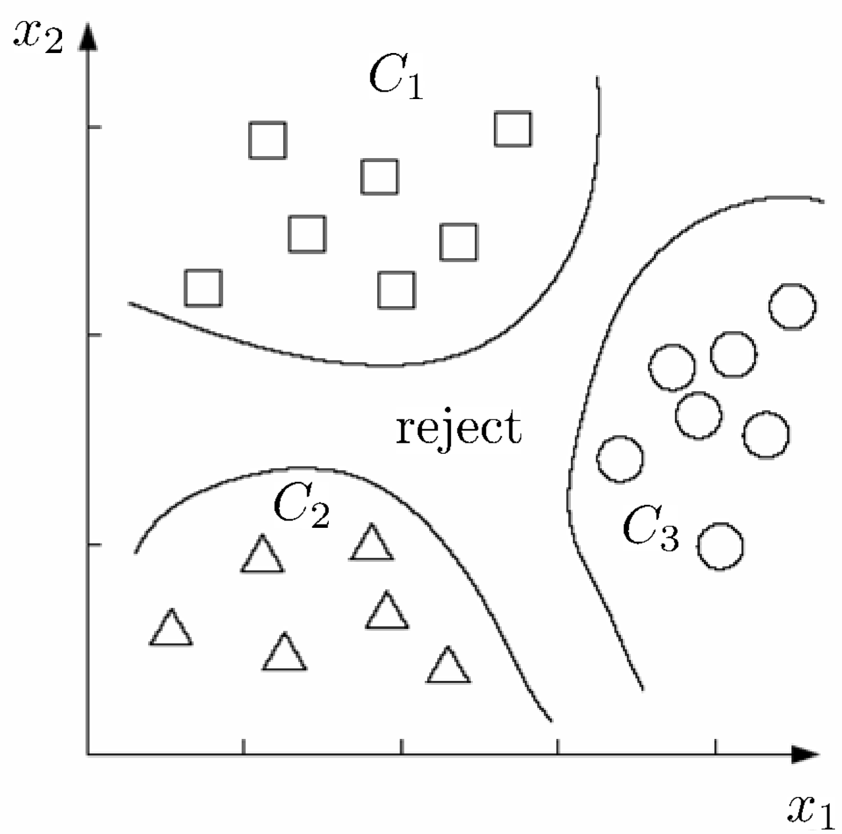
- For each class we create a discriminant function $g_\ell(\mathbf{x})$,

  The classifier assigns class C_ℓ if

$$g_\ell(\mathbf{x}) = \max_k g_k(\mathbf{x})$$

- E.g. 1: $g_\ell(\mathbf{x}) = -R(\alpha_\ell \mid \mathbf{x})$
- E.g. 2a: $g_\ell(\mathbf{x}) = P(C_\ell \mid \mathbf{x})$ or
- E.g. 2b: $g_\ell(\mathbf{x}) = p(\mathbf{x} \mid C_\ell)P(C_\ell)$

Decision Regions



- The discriminant functions define decision regions $R_1, \ldots, R_k$, s.t.,

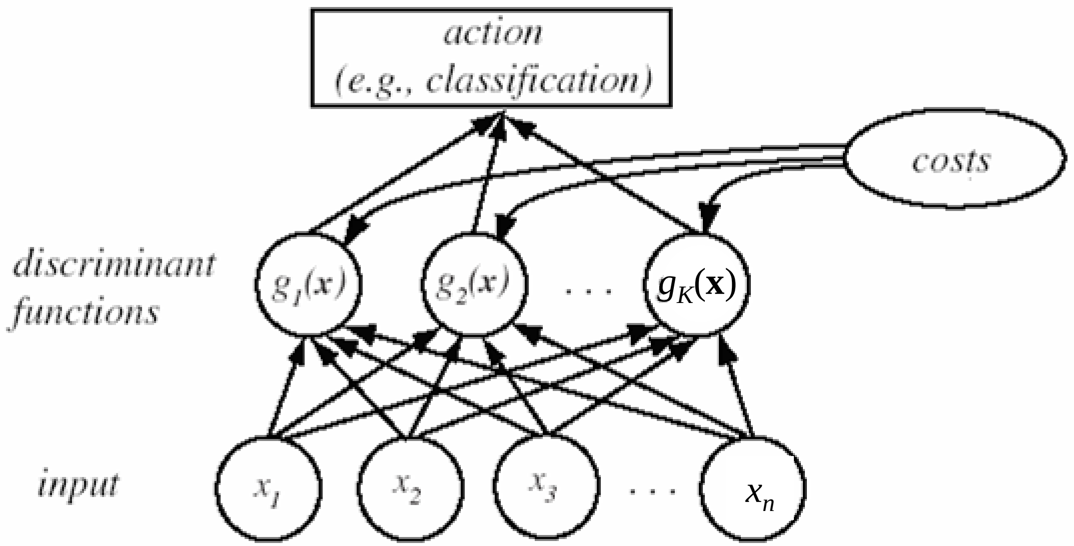
$$R_\ell = \{\mathbf{x} \mid g_\ell(\mathbf{x}) = \max_k g_k(\mathbf{x})\}$$

- When only two classes, we need only one discriminant

$$g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x})$$

choose C_1 if $g(\mathbf{x}) > 0$ and C_2 otherwise

Neural Networks like Classifier



- The functional structure of a general statistical pattern classifier which includes n inputs and K discriminant functions $g_\ell(\mathbf{x})$. A subsequent step determines which of the discriminant values is the maximum, and categorized the input pattern accordingly.

The Relationship to Other Machine Learning Methods

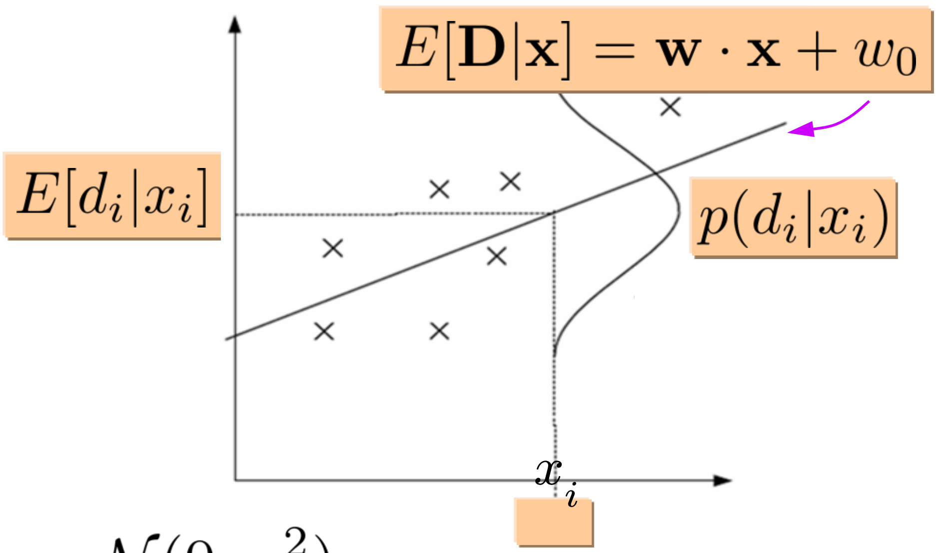
— Bayesian Theory and Bayesian Modeling

MLE, MAP to Explain Other Learning Theories

- Up to now, ML (or MAP) is used as a criterion to select the hypothesis.
  - Parametric methods, or enumeration methods can be applied!
- There are links between ML (or MAP) and other learning theories!
  - Will be shown: under certain assumptions, any Mean Squared Estimation learning algorithm outputs an ML hypothesis
  - MDL and MAP (ML) are related criteria!

ML & MSE hypotheses

- Under certain assumptions any learning algorithm that minimizes the squared error between the output hypothesis predictions and the training data will output a maximum likelihood (ML) hypothesis
  - Least-squared error is also called mean squared error (MSE)



- Problem setting:

$$y_i = h(\mathbf{x}_i) + e_i$$

$$h(\mathbf{x}_i) : \text{noise-free}$$

$$e_i : \text{independently drawn from } \mathcal{N}(0, \sigma^2)$$

Link between ML & MSE

$$\begin{aligned}
h_{ML} &= \operatorname{argmax}_{h \in \mathcal{H}} p(\mathcal{D} \mid h) = \operatorname{argmax}_{h \in \mathcal{H}} \prod_{i=1}^{m} p(y_i \mid h) \\
&= \operatorname{argmax}_{h \in \mathcal{H}} \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - h(\mathbf{x}_i))^2\right) \\
&= \operatorname{argmax}_{h \in \mathcal{H}} \sum_{i=1}^{m} \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}(y_i - h(\mathbf{x}_i))^2 \\
&= \operatorname{argmax}_{h \in \mathcal{H}} \sum_{i=1}^{m} -\frac{1}{2\sigma^2}(y_i - h(\mathbf{x}_i))^2 \\
&= \operatorname{argmin}_{h \in \mathcal{H}} \sum_{i=1}^{m} (y_i - h(\mathbf{x}_i))^2
\end{aligned}$$

- Assuming (1) fixed $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m)$, and (2) the training examples are mutually independent given h

Is noise *ND* distributed?

- Would be nice: easier to analyze mathematically
- Is likely
  - ND approximates other distributions
  - *Central Limit Theorem*: For identically distributed random variables $Y_1, Y_2, \ldots, Y_n$ governed by an arbitrary probability distribution with mean μ and finite variance σ^2.
    The sample mean of them
    $\overline{Y} = \sum_{i=1}^{n} Y_i$ goes to $ND(\mu, \sigma^2/n)$
- CLT applies?
  - Noise = outcome of independent random events?
  - Identically distributed?
- Note: noise only in y_i not in $\mathbf{x}_i$

Other than ML & MAP:
Occam's Razor, MDL & All That

- We talk about Model Selection!
- Occam's Razor (1285 – 1349): "One should not increase, beyond what is necessary, the number of entities required to explain anything."
- **Definition** Minimum Description Length: "Select the hypothesis which minimizes the sum of the length of the description of the hypothesis (also called "model") and the length of the description of the data relative to the hypothesis." or best theory

$$h_{\mathrm{MDL}} = \mathrm{argmin}_h(\mathrm{theory} + \mathrm{exceptions})$$

Occam's Razor

Why prefer short hypotheses?

- Argument in favor:
  - Fewer short hypotheses than long hypotheses
  - A short hypothesis that fits the data is unlikely to be a coincidence;

    a long hypothesis that fits the data might be a coincidence
- Argument opposed:
  - There are many ways to define small sets of hypotheses (notion of coding $\mathrm{length}(x) = -\lg P(x)$, Minimum Description Length…)
  - What is so special about small sets based on *size* of hypothesis

Shannon's Entropy

- Given 5 letter alphabet $\langle A, B, C, D, E \rangle$, how to efficiently encode an article with only these 5 letters?
- Frequencies:

$$P(A) = \frac{1}{4}, P(B) = \frac{1}{16}, P(C) = \frac{1}{8}, P(D) = \frac{1}{16}, P(E) = \frac{1}{2}$$

- One of the optimal codes is:

  A: 10

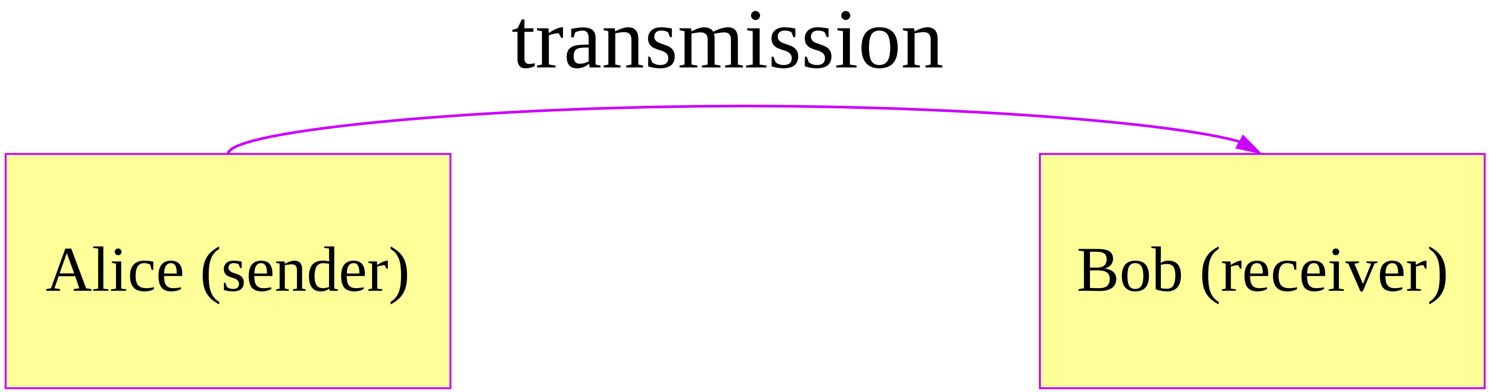
  B: 1110

  C: 110

  D: 1111

  E: 0
- In the optimal code, $\mathrm{length}(x) = -\lg P(x)$

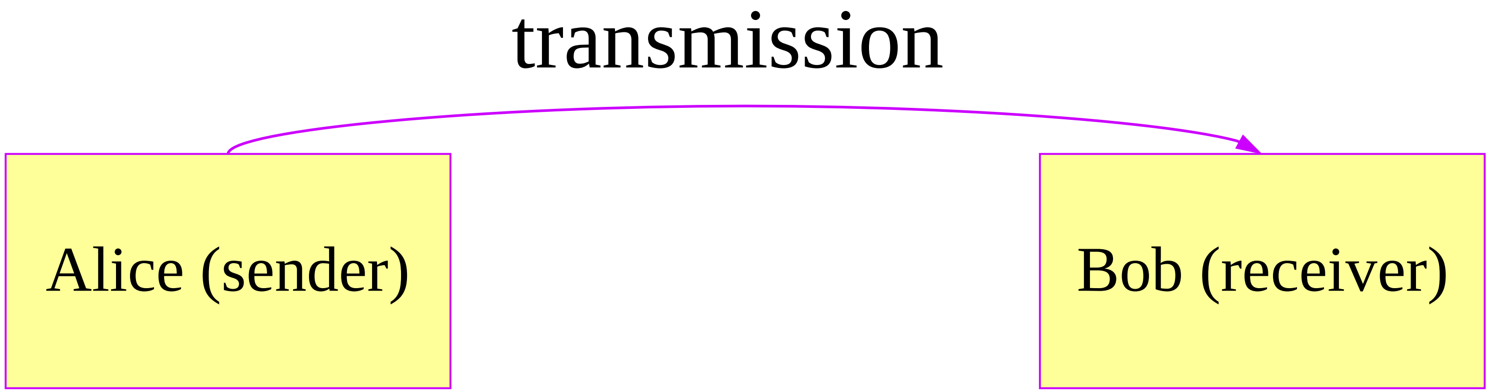
When Transmitting a Set of Symbols



- Assume we want to transmit y_i

  What is the most efficient coding scheme?
- We can adopt Shannon-Fano code, based on the distribution of y_i !

How about Transmitting a Set of Symbols y_i, Given $\mathbf{x}_i$?



- Assume both of sender and receiver know $\mathbf{x}_i$,we want to transmit y_i
- What is the most efficient coding scheme?
- In general, we can take advantage of the relation between $\mathbf{x}_i$ and y_i!

Transmission and Shannon-Fano code

- Case: finding the most economic way to transmit a data set $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^m$ to the receiver side, assuming the receiver knowing $\{\mathbf{x}_i\}_{i=1}^m$ already?
- Shannon-Fano code: assigning prefix code length $\ell_x := -\lg P(x)$ to the symbol x of probability $P(x)$
- Idea 1: assigning Shannon-Fano code based on the $\{y_i\}_{i=1}^m$ distribution of
- Idea 2: finding a theory to describe the relation between $\{\mathbf{x}_i\}_{i=1}^m$ and $\{y_i\}_{i=1}^m$, then transmitting the theory first, followed by some exceptions

An Example

- Data: $\{(x,y) = (1,3), (0,-1), (-4,3), (-3,-1), (2,\boxed{3}), (-1,\boxed{0})\}$
- A theory: $y = x^2 + 3x - 1$?

  Perfection: $\{(1,3), (0,-1), (-4,3), (-3,-1), (2,\boxed{9}), (-1,\boxed{-3})\}$

  Exceptions: $\{(2,3), (-1,0)\}$
- Transmission:

  theory $\rightarrow (2,1,1,3,0,-1)$

  Basically we assign short codes for simple functions

  corrections $\rightarrow \{0,0,0,0,-6,3\}$, with uncertainty

  Because $P(0,0,0,0,-6,3)$ is high and it should be encoded by a short code

Information Theory & MDL

- Interpretation of h_{MAP} by information theory

$$\begin{aligned} & \operatorname{argmax}_h P(\mathcal{D} \mid h) P(h) \\ = & \operatorname{argmax}_h \lg P(\mathcal{D} \mid h) + \lg P(h) \\ = & \operatorname{argmin}_h - \lg P(h) - \lg P(\mathcal{D} \mid h) \\ = & \text{“short hypotheses preferred”?} \end{aligned}$$

- IT: length-optimal coding
  - symbol i has probability p_i
  - fact: optimality is reached by $-\lg p_i$ bits for i

Link between MAP, ML and MDL

- Description lengths

$L_C(i)$: length of coding i in C

$C_{\mathcal{H}}$: optimal code for $\mathcal{H}$

$C_{\mathcal{D}|h}$: optimal code for $\mathcal{D}$ given h

$-\log P(h)$: code length of h when h is in the optimal coding for $\mathcal{H}$

$-\log P(\mathcal{D} \mid h)$: code length of the optimal coding for $\mathcal{D}$ given h

- $$\begin{aligned} h_{\mathrm{MAP}} &= \operatorname{argmin}_h L_{C_{\mathcal{H}}}(h) + L_{C_{\mathcal{D}|h}}(\mathcal{D} \mid h) \quad (\text{given } \mathcal{H}, \mathcal{D}) \\ h_{\mathrm{ML}} &= \operatorname{argmin}_h L_{C_{\mathcal{D}|h}}(\mathcal{D} \mid h) \quad (\text{given } \mathcal{H}, \mathcal{D}) \end{aligned}$$

(dangerous of overfitting, check coin tossing!)

$$h_{\mathrm{MDL}} = \operatorname{argmin}_h L_{C_1}(h) + L_{C_2}(\mathcal{D} \mid h) \quad (\text{given } C_1, C_2)$$

MDL Reconsidered

- Illustrates the tradeoff between
  - accuracy of hypothesis and
  - length of hypothesis
  - short with few errors vs. perfect long
- IF $L_{C_1}(h) = -\log P(h)$ (the particular optimal code)
  AND $L_{C_2}(\mathcal{D} \mid h) = -\log P(\mathcal{D} \mid h)$ (the particular optimal code)
  THEN $h_{\text{MDL}} = h_{\text{MAP}}$
- Note 1: no reason to believe that MDL + *arbitrary* C_1, C_2 is better
- Note 2: for a successful MDL, we need
  1. appropriate hypothesis space,
  2. correct density estimation and
  3. good coding scheme

References

- “Machine Learning”, T. M. Mitchell, 1997
- A. Hertzmann’s SIGGRAPH 2004 Tutorial on Bayesian Learning
- Bayesian Decision Theory:

  “Pattern Classification” (2nd ed) by R. O. Duda, P. E. Hart and D. G. Stork, John Wiley & Sons, 2000 (Ch. 2)