NB: Notes based
Alpaydin chp 3
Bayesian Decision Theory

Probability & Inference

- Result of tossing a coin is $\in \{H, T\}$

- Random variable $X \in \{1, 0\}$

  - Recall the Bernoulli PMF

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

- Sample: $X = \{x^t\}_{t=1}^{N}$

- Estimation (MLE): $\hat{\theta} = \dfrac{\# H}{\# \text{Tosses}} = \dfrac{1}{N} \sum_{t=1}^{N} x^t$

- Predict the next toss:

  H if $\hat{\theta} > \frac{1}{2}$, T otherwise

Conjugate Bayes

$\theta \sim \text{Beta}(a, b)$

$\theta | X \sim \text{Beta}(a^*, b^*)$

$a^* = a + N_1$

$b^* = b + N_0$

where $N_1 = \#$ heads
$N_0 = \#$ tails
$N_0 + N_1 = N$ # of trials

Classification

- Credit scoring: Inputs are income and savings
  - Output: low-risk vs high-risk

- Input: $X = [x_1, x_2]^T$

  Output: $C \in \{0, 1\}$

- Prediction

  choose/select $\begin{cases} C = 1 & \text{if } P(C=1 \mid x_1, x_2) > 0.5 \\ C = 0 & \text{otherwise} \end{cases}$

  or

  choose/select $\begin{cases} C = 1 & \text{if } P(C=1 \mid x_1, x_2) > P(C=0 \mid x_1, x_2) \\ C = 0 & \text{otherwise} \end{cases}$

Bayes' Rule

NB:
$P(C=0|x) + P(C=1|x) = 1$

Posterior, likelihood, Prior

$$\underbrace{P(C|x)}_{\text{Posterior}} = \frac{\overbrace{P(x|C)}^{\text{likelihood}}\ \overbrace{P(C)}^{\text{Prior}}}{P(x)}$$

$P(x)$ ← the "evidence" or marginal of the data

Decision

choose $C=1$ if

$P(C=1|x) > P(C=0|x)$

Prior prob. $P(C=1)$: prob. of high-risk customer

- $P(C=0) + P(C=1) = 1$

Class Likelihood : $P(x|C)$ prob. of event belonging to class C given the observable x

(Q: how ~~likely~~ likely are the data if the true state of the world is C)

- $P(x_1, x_2 | C)$ = prob. that high-risk customer will have $[x_1, x_2]^T$

Evidence : $P(x)$: Marginal prob. of $[x_1, x_2]^T$

$P(x) = P(x|C=1)P(C=1) + P(x|C=0)P(C=0)$

Bayes' Rule for $K > 2$ Classes (Mutually exclusive & exhaustive)

Prior Probabilities : $P(C_i) \quad \forall\, i = 1, \ldots, K$

$$\sum_{i=1}^{K} P(C_i) = 1$$

Class Likelihood : $P(x \mid C_i)$

Posterior prob

$$P(C_i \mid x) = \frac{P(x \mid C_i)\, P(C_i)}{P(x)}$$

$$= \frac{P(x \mid C_i)\, P(C_i)}{\sum_{k=1}^{K} P(x \mid C_k)\, P(C_k)}$$

Decision:

Choose as the output the class C_i that has the maximux posterior prob.

$$P(C_i \mid x)$$

Losses and Risks

- Credit scoring decisions should be made so as to maximize gains / limit losses

- Classes $C_1, \dots, C_K$

- Let α_i = decision to assign C_i to the input $\forall\ 1 \le i \le k$

- Let λ_{ik} = the loss from assigning C_i to input that belongs to C_k (i.e. misclassification)

- Expected risk for α_i:

$$R(\alpha_i \mid x) = \sum_{k=1}^{K} \lambda_{ik} P(C_k \mid x)$$

- Decision: choose α_i with minimal expected risk

$$R(\alpha_i \mid x) = \min_k R(\alpha_k \mid x)$$

0/1 Loss Case

- Correct decisions have ∅ loss
  Incorrect decisions have 1 loss

NB:
- all equal
- losses are symmetric

$$\lambda_{ik} = \begin{cases} \emptyset & \text{if } i = k \\ 1 & \text{if } i \neq k \end{cases}$$

- The risk of taking action α_i is:

$$R(\alpha_i \mid x) = \sum_{k=1}^{K} \underbrace{\lambda_{ik}}_{\emptyset \text{ or } 1} P(C_k \mid x) = \sum_{k \neq i}^{K} P(C_k \mid x) = \underbrace{1 - P(C_i \mid x)}_{\text{via compliment rule}}$$

- **Decision:** to minimize risk, assign C_i to the most probable class

Losses & Risks: Reject

NB:
- reject option might require human ~~[illegible]~~ analysts
- or another ML algorithm

- Misclassification: very high cost

- Consider an extra option $(K+1)$-st class "reject"

- Loss function:

$$\lambda_{ik} = \begin{cases} \emptyset & \text{if } i = k \\ \lambda & \text{if } i = K+1, \quad 0 < \lambda < 1 \\ 1 & \text{if } i \neq k \text{ and } i = k+1 \end{cases}$$

- Risk of reject:

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

NB:

$$\lambda \underbrace{\sum_{k=1}^{K} P(C_k|x)}_{1}$$

- Risk of misclassification:

$$R(\alpha_i | x) = \sum_{k \neq i} P(C_k | x) = \underbrace{1 - P(C_i | x)}_{\text{via compliment rule}}$$

(i.e. one minus "doing it correctly")

Decision:
- Choose C_i $\min_{1 \le i \le k} R(\alpha_i | x)$

output:
- C_i if $P(C_i|x) > 1 - \lambda$ ($P(C_i|x) > P(C_k|x)$)
- Reject otherwise

Cases:

$\lambda = 0$ "always reject" (as good as correct classification)

$\lambda = 1$ "never reject" (as bad as incorrect classification)

Discriminant Functions

- Choose $g_i(x)$, $i = 1, \dots, k$ s.t. output C_i if $g_i(x) = \max_k g_k(x)$

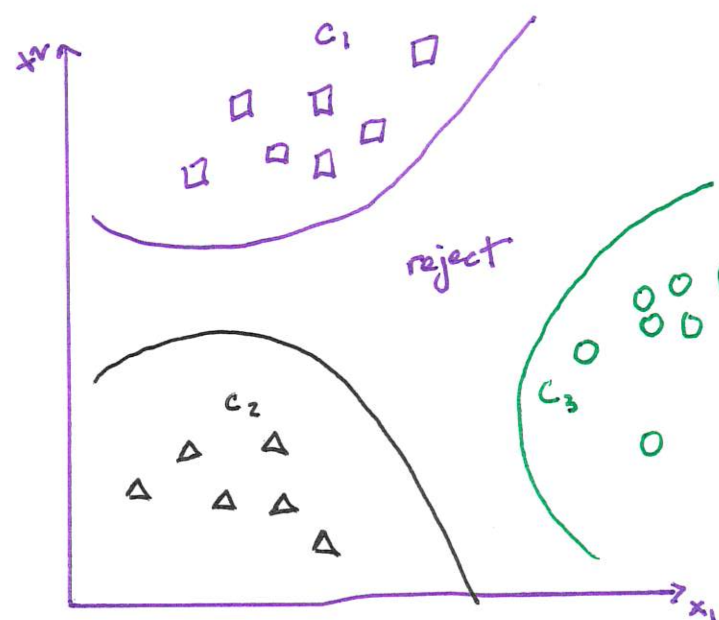
- <u>Bayes' Classifier</u>: $g_i(x) = -R(\alpha_i|x)$

  - 0/1 loss: $g_i(x) = P(C_i|x) = \dfrac{P(x|C_i)P(C_i)}{P(x)}$ via Bayes' Rule

    Equivalent way: $g_i(x) = P(x|C_i)P(C_i)$ (simpler/easier)

- k decision regions

$$R_i = \{ x \mid g_i(x) = g_k(x) \}$$



$K = 2$ Classes

- Dichotomizer $(K=2)$ vs Polychotomizer $(K>2)$
- $g(x) = g_1(x) - g_2(x)$

$$\text{choose} \begin{cases} C_1 & \text{if } g(x) > 0 \\ C_2 & \text{otherwise} \end{cases}$$

- Log odds:

$$\log \frac{P(C_1 \mid x)}{P(C_2 \mid x)} = \underbrace{\log P(C_1 \mid x) - \log P(C_2 \mid x)}_{\text{log difference in probabilities}}$$

Bayesian Thinking Revisited

NB: notes based on chp. 3 Losses and Decision Making
An Introduction to Bayesian Thinking

Introduction to Bayes Factors

- Return to HIV testing

  - H_1: Patient does not have HIV
  - H_2: Patient does have HIV

  - Actions / Decisions

    - d_1: choose H_1
    - d_2: choose H_2

  - $L(d)$ the loss associated w/ decision d_i $\forall i \in \{1,2\}$

    - $d = d_1$
      - Right: decide H_1 / and they don't $\Rightarrow L(d_1) = \emptyset$
      - Wrong: decide H_1 / and they do $\Rightarrow L(d_1) = \omega_1$

    - $d = d_2$
      - Right: decide H_2 / they do $\Rightarrow L(d_2) = \emptyset$
      - Wrong: decide H_2 / they don't $\Rightarrow L(d_2) = \omega_2$

Losses

$$L(d_1) = \begin{cases} \emptyset & \text{if } d_1 \text{ correct} \\ w_1 = 1000 & \text{else} \end{cases}$$

$$L(d_2) = \begin{cases} \emptyset & \text{if } d_2 \text{ correct} \\ w_2 = 10 & \text{else} \end{cases}$$

H_1: Patient does not ~~has~~ HIV

H_2: Patient has HIV

Posteriors

- (+) stands for a positive result from the ELISA
- $P(H_1 \mid +) \approx 0.88$ posterior prob. of NOT having HIV given a (+) ELISA result
- $P(H_2 \mid +) \approx 0.12$ Posterior prob. of Having HIV given the (+) ELISA result (from the compliment)

Expected Losses

- $E(L(d_1)) = 0.88(\phi) + 0.12(1000) = 120$

  $E[L(d_2)] = 0.88(10) + 0.12(\phi) = 8.8$

- Since $E[L(d_2)] < E[L(d_1)]$ $\Rightarrow$ decide the patient has HIV

NB:

- Decision highly influenced by losses assigned to d_1 and d_2

- If losses symmetric, say $\omega_1 = \omega_2 = 10$
  - $E[L(d_1)] = 0.88(\phi) + 0.12(10) = 1.2$
  - while $E[L(d_2)]$ would not change
  - We would decide that Patient does NOT have HIV!

Bayes factors

- Continue with HIV testing example

- __Prior odds__ = ratio of the prior probabilities of hypotheses

$$O[H_1 : H_2] = \frac{P(H_1)}{P(H_2)}$$

- __Posterior odds__ = ratio of the two posterior probabilities of hypotheses

$$PO[H_1 : H_2] = \frac{P(H_1 \mid Data)}{P(H_2 \mid Data)}$$

Using Bayes' Rule, we find that

$$PO[H_1 : H_2] = \frac{P(H_1 \mid Data)}{P(H_2 \mid Data)}$$

$$= \frac{P(data \mid H_1)\, P(H_1) \,/\, P(data)}{P(data \mid H_2)\, P(H_2) \,/\, P(data)}$$

$$= \frac{P(data \mid H_1)\; P(H_1)}{P(data \mid H_2)\; P(H_2)}$$

$$= \underbrace{\frac{P(data \mid H_1)}{P(data \mid H_2)}}_{\text{Bayes Factor}} * \underbrace{\frac{P(H_1)}{P(H_2)}}_{\text{Prior odds}}$$

So, we have that

$$PO[H_1 : H_2] = BF[H_1 : H_2] * O[H_1 : H_2]$$

- BF quantifies the evidence of data arising from H_1 versus H_2

- In the discrete case: $BF[H_1 : H_2] = \dfrac{P(\text{data} \mid H_1)}{P(\text{data} \mid H_2)}$

  (likelihood ratio)

- In the continuous case

$$BF[H_1 : H_2] = \frac{\int P(\text{data} \mid \theta, H_1)\, d\theta}{\int P(\text{data} \mid \theta, H_2)\, d\theta}$$

NB: θ is the index of all possible models/parameters

- Continue w/ HIV case

  H_1: Patient does not have HIV

  H_2: Patient does have HIV

- Priors:

  $P(H_1) = .99852$

  $P(H_2) = .00148$

  NB: from prevalence of HIV in the population

- Prior odds then is

$$O[H_1 : H_2] = \frac{P(H_1)}{P(H_2)} = \frac{.99852}{.00148} = 674.6757$$

- Posteriors

$$P(H_1 \mid +) = .8788551$$

$$P(H_2 \mid +) = .1211449$$

$$PO[H_1 : H_2] = \frac{.8788551}{.1211449} = 7.254578$$

- Bayes Factor

$$BF[H_1 : H_2] = \frac{PO[H_1 : H_2]}{O[H_1 : H_2]}$$

$$= \frac{7.25457}{674.6757} = 0.0108$$

$$= \frac{P(+ \mid H_1)}{P(+ \mid H_2)} = \frac{.01}{.93} \approx .0108$$

- So now that we have calculated $BF[H_1 : H_2]$, how should understand it's meaning?

- Jeffrey (1961)

$BF[H_1 : H_2]$	Evidence against H_2
1 to 3	Not worth a bare mention
3 to 20	Positive
20 to 150	Strong
> 150	Very stong

$2 * \log(BF[H_1 : H_2])$	Evidence against H_1
0 to 2	Not worth a bare mention
2 to 6	Positive
6 to 10	Strong
> 10	Very strong

- NB: notice that for the HN case it doesn't even appear on the scale.
- Let's reorder so that it's $BF[H_2 : H_1] = \frac{1}{BF[H_1 : H_2]} = \frac{1}{.0108} = \underline{\underline{92.59259}}$

$$\approx 93$$

- Hence the evidence against H_1 is ~~[illegible]~~ "Strong"