# CSE 569 - Fundamentals of Statistical Learning: Homework #1

Due on September 22, 2017 at 11:59pm

*Dr. Baoxin Li*

**Andrew Dudley**

# Problem 1

Consider the following decision rule for a two-category, one-dimensional problem:

$$\text{Decide } \omega_1 \text{ if } x > 0; \text{ otherwise, decide } \omega_2$$

a) Show that the probabilty of error for this rule is given by

$$P(\text{error}) = P(\omega_1)\int_{-\infty}^{\theta} P(x|\omega_1)dx + P(\omega_2)\int_{\theta}^{\infty} P(x|\omega_2)dx$$

**Solution**
Given the decision rule, we know that there are two cases where we will incur error:

(1) $x < \theta$ when the true state of nature is $\omega_1$
(2) $x \geq \theta$ when the true state of nature is $\omega_2$

Define $R_1 = \{x \,|\, x > \theta\}$ and $R_2 = \{x \,|\, x \leq \theta\}$. The probability of error for the decision rule can then be written as

$$\begin{aligned}
P(error) &= P(x \in R_2, \omega_1) + P(x \in R_1, \omega_2) \\
&= P(x \in R_2|\omega_1)P(\omega_1) + P(x \in R_1|\omega_2)P(\omega_2) \\
&= \int_{R_2} P(x|\omega_1)P(\omega_1)dx + \int_{R_1} P(x|\omega_2)P(\omega_2)dx \\
&= P(\omega_1)\int_{R_2} P(x|\omega_1)dx + P(\omega_2)\int_{R_1} P(x|\omega_2)dx \\
&\boldsymbol{= P(\omega_1)\int_{-\infty}^{\theta} P(x|\omega_1)dx + P(\omega_2)\int_{\theta}^{\infty} P(x|\omega_2)dx}
\end{aligned}$$

b) By differentiating, show that a necessary condition to minimize P(error) is that $\theta$ satisfy

$$P(\theta \,|\, \omega_1)P(\omega_1) = p(\theta \,|\, \omega_2)P(\omega_2)$$

**Solution**
Using Leibniz' Integral Rule,

$$\frac{d}{d\theta}\left[P(\omega_1)\int_{-\infty}^{\theta} P(x|\omega_1)dx + P(\omega_2)\int_{\theta}^{\infty} P(x|\omega_2)dx\right]$$

$$\begin{aligned}
&= P(\omega_1)\left[\int_{a}^{\infty} \frac{\delta}{\delta\theta}[P(x \,|\, \omega_1)]dx + P(\omega \,|\, \omega_1)\cdot\frac{d}{d\theta}\theta - P(a \,|\, \omega_1)\cdot\frac{d}{d\theta}a\right] \\
&\quad + P(\omega_2)\left[\int_{\theta}^{b} \frac{\delta}{\delta\theta}[P(x \,|\, \omega_2)]dx + P(\omega \,|\, \omega_1)\cdot\frac{d}{d\theta}\theta - P(a \,|\, \omega_1)\cdot\frac{d}{d\theta}a\right] \\
&= P(\theta \,|\, \omega_1)P(\omega_1) - p(\theta \,|\, \omega_2)P(\omega_2)
\end{aligned}$$

To minimize P(error), we will need a mininum value of the function, and the rate of change of P(error) at that point will be 0.

$$P(\theta \,|\, \omega_1)P(\omega_1) - p(\theta \,|\, \omega_2)P(\omega_2) = 0$$

which gives us

$$\mathbf{P}(\theta \,|\, \boldsymbol{\omega_1})\mathbf{P}(\boldsymbol{\omega_1}) = \mathbf{p}(\theta \,|\, \boldsymbol{\omega_2})\mathbf{P}(\boldsymbol{\omega_2}).$$

Problem 1 continued on next page. . .

c) Does this equation satisfy $\theta$ uniquely?
**Solution**
No. If the class conditional probability density functions intersect more than once - resulting in disjoint regions - then there will be multiple values of theta that satisfy the criterion

$$P(\theta \,|\, \omega_1)P(\omega_1) - p(\theta \,|\, \omega_2)P(\omega_2) = 0$$

d) Give an example where a value of $\theta$ satisfying the equation actually maximizes the probability of error.
**Solution**
If the class-conditional probability density distributions for $\omega_1$ and $\omega_2$ are Gaussian and the mean of the gaussian distribution for $\omega_1$ is less than the mean for $\omega_2$, then $\theta$ will maximize the probability of error.

# Problem 2

Consider a one-dimensional, two-category classification problem, with equal prior probabilities ($P(\omega_1) = P(\omega_2) = \frac{1}{2}$). The class-conditional PDFs for the two classes are the normal densities $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$, respectively. Note that these two PDFs have the same variance $\sigma^2$. $N(\mu, \sigma^2)$ denotes the normal density defines by

$$p(x \,|\, \mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

We further assume that the losses $\lambda_{11} = \lambda_{22} = 0$, but $\lambda_{12}$ and $\lambda_{21}$ are non-zero. Find the optimal decision rule for classifying any feature point x.

**Solution**

In general, a two-category classification problem reduces to the Bayes decision rule:

$$\text{Decide } \omega_1 \text{ if } (\lambda_{21} - \lambda_{11})p(x \,|\, \omega_1)P(\omega_1) > (\lambda_{12} - \lambda_{22})p(x \,|\, \omega_2)P(\omega_2)\text{; otherwise, decide } \omega_2$$

Because of the equal variance and equal prior probabilities, the equation of the decision rule reduces to

$$\begin{aligned} ln(\lambda_{21}) + (-[x^2 - 2x\mu_1 + \mu_1^2]) &> ln(\lambda_{12}) + (-[x^2 - 2x\mu_2 + \mu_2^2]) \\ \implies \qquad 2x\mu_1 - 2x\mu_2 &> -\mu_2^2 + \mu_1^2 + ln\left(\frac{\lambda_{12}}{\lambda_{21}}\right) \end{aligned}$$

We can define the region $R_1$ by solving for x

$$\mathbf{R_1} = \left\{ \mathbf{x} \;\middle|\; \mathbf{x} < \frac{\boldsymbol{\mu_2^2} + \boldsymbol{\mu_1^2} + \mathbf{ln}\left(\frac{\boldsymbol{\lambda_{12}}}{\boldsymbol{\lambda_{21}}}\right)}{\mathbf{2}(\boldsymbol{\mu_2} - \boldsymbol{\mu_1})} \right\}$$

Which results in the optimal decision boundary

$$\textbf{Decide } \boldsymbol{\omega_1} \text{ if } \mathbf{x} \in \mathbf{R_1}\textbf{; otherwise, decide } \boldsymbol{\omega_2}$$

# Problem 3

Consider a two-class classification problem with one-dimensional class-conditionals given as

$$P(x \,|\, \omega_1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-1)^2}{2}},$$

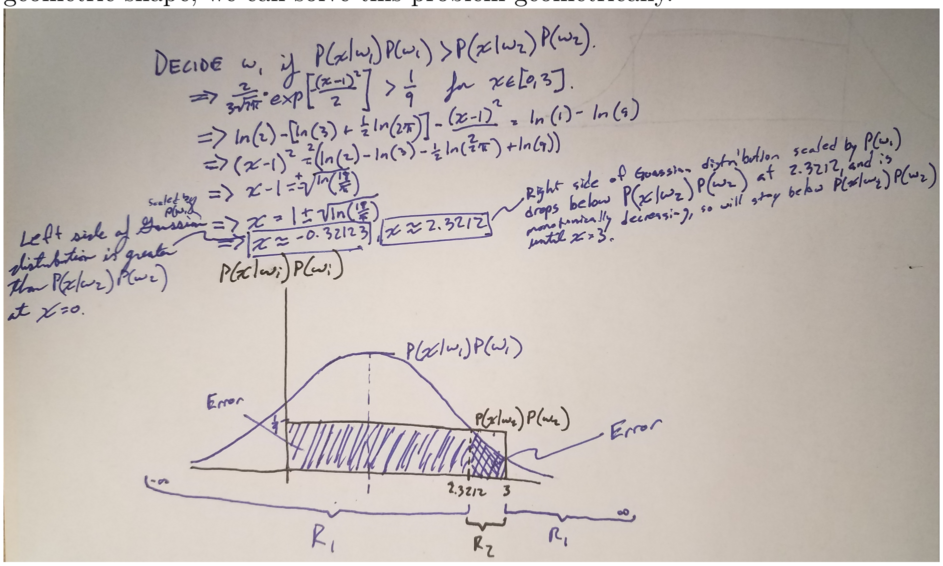
Problem 3 continued on next page. . . 

$$P(x \mid \omega_2) = \begin{cases} \frac{1}{3}, & \text{if } 0 \leq x \leq 3 \\ 0, & \text{otherwise} \end{cases}$$

Suppose the priors are $P(\omega_1) = \frac{2}{3}, P(\omega_2) = \frac{1}{3}$. Find the optimal decision rule for doing the classification. What is the Bayes error in this case?

**Solution**

Given that the class-conditional PDF for $\omega_1$ is a Guassian distribution and the PDF for $\omega_2$ is a simple geometric shape, we can solve this problem geometrically.




$P(x|\omega_1)P(\omega_1)$ (the posterior probability scaled by $p(x)$) intersects the line $y = 1/9$ at two points, once at $x = -0.32123$ (which outside of the domain of $P(x|\omega_2)$) and once at $x \approx 2.3212$. This tells us that the region where $P(\omega_1 \mid x) > P(\omega_2 \mid x)$ can be defined as

$$\mathbf{R_1} = \{\mathbf{x} \mid \mathbf{x} < \mathbf{2.3212} \text{ or } \mathbf{x} > \mathbf{3}\}$$

This problem has a zero-one loss function, so the conditional risk can be calculated as

$$R(\alpha_i \mid x) = 1 - P(\omega_i \mid x)$$

and the Bayes Rules to minimize risk says to minimize the conditional risk, so we choose the value of $i$ that maximizes $P(\omega_i \mid x)$, leading us to the decision rule

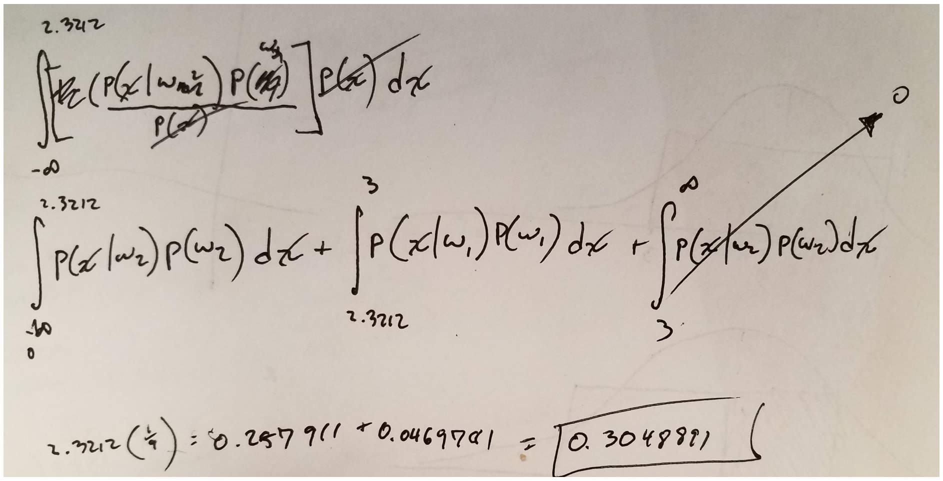
**Decide $\omega_1$ if $\mathbf{x} \in \mathbf{R_1}$; otherwise, decide $\omega_2$**

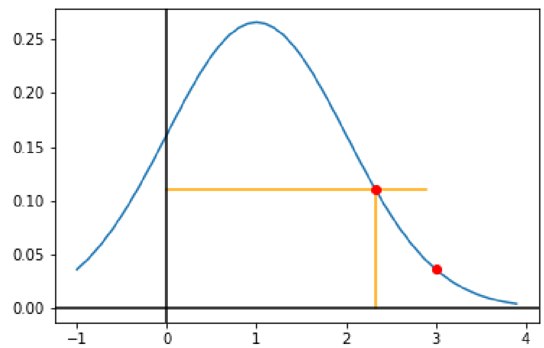
The Bayes error can be calculated with the formula

$$P(error) = P(\omega_1) \int_{R_2} P(x|\omega_1)dx + P(\omega_2) \int_{R_1} P(x|\omega_2)dx$$

Worked out below...

Problem 3 continued on next page...




Therefore, $\mathbf{P(error)} \approx \mathbf{0.30489}$

# Problem 4

Consider the following Bayesian network where all of the nodes are assigned to be binary random variables

a) Suppose that X is measured and its value is $x_1$, compute the probability that we will observe W having a value $\omega_1$. i.e., $P(\omega_0 \,|\, x_1)$

**Solution**

Need the marginalized probability of $\omega_0$ conditioned on $x_1$.

$$\begin{aligned} P(\omega_0 \,| x_1) &= p(x_1)\frac{\sum_Y \sum_Z P(Y \,|\, x_1)P(Z \,|\, Y)P(w_0 \,|\, Z)}{p(x_1)} \\ &= \sum_Y P(Y \,|\, x_1) \sum_Z P(Z \,|\, Y)P(w_0 \,|\, Z) \\ &= 0.6(0.4(0.7) + 0.6(0.55)) + 0.4(0.75(0.7) + 0.25(0.55)) \\ &= \mathbf{0.631} \end{aligned}$$

b) Suppose that W is measured and its value is $w_1$, compute the probability that we will observe X having a value $x_0$.

$$\begin{aligned}
P(x_0 \mid w_1) &= \frac{P(x_0, w_1)}{P(w_1)} \\
&= \frac{\sum_Y \sum_Z P(x_0)P(Y \mid x_0)P(Z \mid Y)P(w_1 \mid Z)}{\sum_X \sum_Y \sum_Z P(X)P(Y \mid X)P(Z \mid Y)P(w_1 \mid Z)} \\
&= \frac{P(x_0) \sum_Y P(Y \mid x_0) \sum_Z P(Z \mid Y)P(w_1 \mid Z)}{\sum_X P(X) \sum_Y P(Y \mid X) \sum_Z P(Z \mid Y)P(w_1 \mid Z)} \\
&= \frac{0.4[0.7(0.4 * 0.3 + 0.6 * 0.45) + 0.3(0.75 * 0.3 + 0.25 * 0.45)]}{\begin{array}{l}0.4[0.7(0.4 * 0.3 + 0.6 * 0.45) + 0.3(0.75 * 0.3 + 0.25 * 0.45)] \\ \quad + 0.6[0.6(0.4 * 0.3 + 0.6 * 0.45) + 0.4(0.75 * 0.3 + 0.25 * 0.45)]\end{array}} \\
&= 0.403395
\end{aligned}$$

# Problem 5

True or False: For a two-class classification problem using the minimum-error-rate rule, in general the decision boundary can take any form. However, if the underlying class-conditionals are Gaussian densities, then the decision boundary is linear (hyperplanes).

**Solution**

False. If the covariance matrices for the Gaussian class-conditional PDFs are different, the disriminant function will be

$$g_i(x) = \mathbf{x^t W_i x} + \mathbf{W_i x} + w_{i0}$$

where $\mathbf{x^t W_i x}$ is inherently quadratic.

# Problem 6

a) Consider the following game: Someone shows you three hats and tells you that there is a prize in one of them. He asks you to choose one of the hats. You choose one hat and tell him which one you chose. He then lifts one of the hats you didn't choose and there is nothing under that hat. He then tells you that you can either stay with the hat you have originally chosen or switch to the other remaining hat. What should you do? Explain your answer.

**Solution**

You should always switch.

Intuitively, after you have chosen a hat there are two possible outcomes: 1/3 of the time, you will choose correctly, in which case the carnie can choose either of the remaining hats to lift. 2/3 of the time, you will choose wrong, in which case the carnie will be forced to lift the hat that doesn't contain the prize (inherently pointing out the hat that *does* have the prize).

**Bayes Rule proof:**

A priori, $P(h_1) = P(h_2) = P(h_3) = 1/3$ where $P(h_i)$ is the prior probability that the prize is under hat i.

Assume we choose hat 1. The carnie knows which hat the prize is under, so we can calculate the class-conditional probability mass functions for the observations $D$ that the carnie chooses to flip hat

Problem 6 continued on next page...

2 or hat 3.

$$P(d_2 \mid h_1) = \frac{1}{2} \qquad P(d_3 \mid h_1) = \frac{1}{2}$$
$$P(d_2 \mid h_2) = 0 \qquad P(d_3 \mid h_2) = 1$$
$$P(d_2 \mid h_3) = 1 \qquad P(d_3 \mid h_3) = 0$$

From the Bayes Rule, we know that

$$P(H \mid D) = \frac{P(D \mid H)P(H)}{\sum_H P(D \mid H)P(H)}$$

So the posteriors can be calculated as:

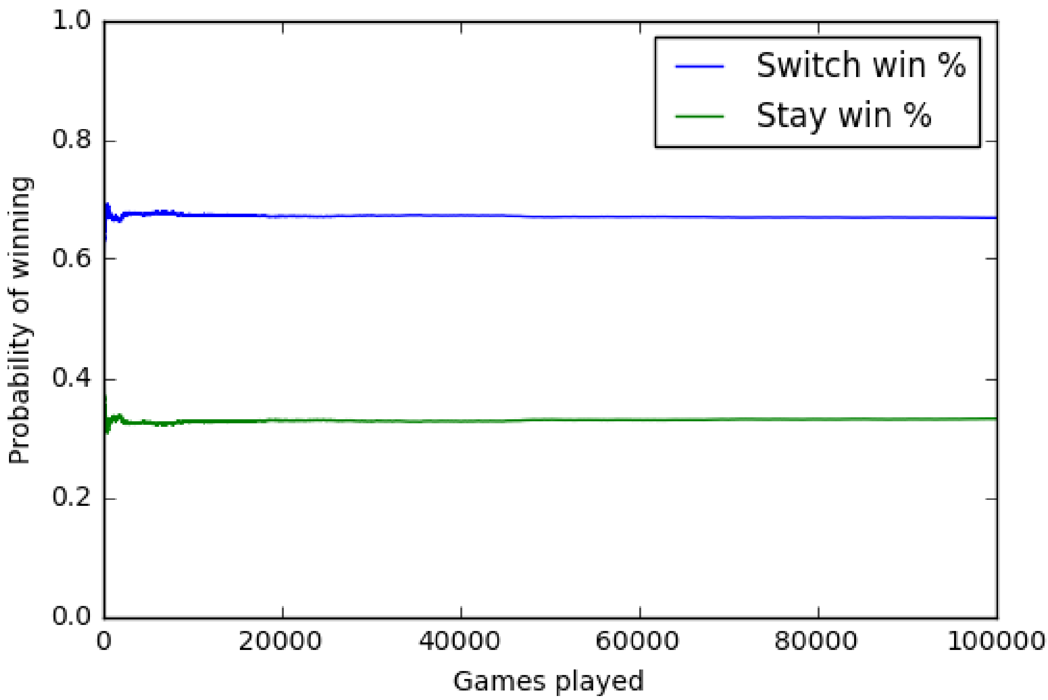
$$P(h_1 \mid d_2) = \frac{P(d_2 \mid h_1)P(h_1)}{\sum_H P(d_2 \mid H)P(H)} = \frac{\frac{1}{2} * \frac{1}{3}}{\frac{1}{2} * \frac{1}{3} + 0 * \frac{1}{3} + 1 * \frac{1}{3}} = \frac{1}{3}$$
$$P(h_2 \mid d_2) = \frac{P(d_2 \mid h_2)P(h_2)}{\sum_H P(d_2 \mid H)P(H)} = \frac{0 * \frac{1}{3}}{\frac{1}{2} * \frac{1}{3} + 0 * \frac{1}{3} + 1 * \frac{1}{3}} = 0$$
$$P(h_3 \mid d_2) = \frac{P(d_2 \mid h_3)P(h_3)}{\sum_H P(d_2 \mid H)P(H)} = \frac{1 * \frac{1}{3}}{\frac{1}{2} * \frac{1}{3} + 0 * \frac{1}{3} + 1 * \frac{1}{3}} = \frac{2}{3}$$

$$P(h_1 \mid d_3) = \frac{P(d_3 \mid h_1)P(h_1)}{\sum_H P(d_3 \mid H)P(H)} = \frac{\frac{1}{2} * \frac{1}{3}}{\frac{1}{2} * \frac{1}{3} + 1 * \frac{1}{3} + 0 * \frac{1}{3}} = \frac{1}{3}$$
$$P(h_2 \mid d_3) = \frac{P(d_3 \mid h_2)P(h_2)}{\sum_H P(d_3 \mid H)P(H)} = \frac{0 * \frac{1}{3}}{\frac{1}{2} * \frac{1}{3} + 1 * \frac{1}{3} + 0 * \frac{1}{3}} = \frac{2}{3}$$
$$P(h_3 \mid d_3) = \frac{P(d_3 \mid h_3)P(h_3)}{\sum_H P(d_3 \mid H)P(H)} = \frac{1 * \frac{1}{3}}{\frac{1}{2} * \frac{1}{3} + 1 * \frac{1}{3} + 0 * \frac{1}{3}} = 0$$

Clearly, regardless of which hat the carnie flips, it is always in our best interest to switch.

b) Design a computer-based experiment to simulate the above game to verify your answer, by playing the game many times to obtain an averaged performance.

**Results**




**Code**

```python
from random import shuffle, randrange
import matplotlib.pyplot as plt

def lprint(text):
    global show_text
    if show_text == True:
        print(text)

# Config
show_text = False

# Counters
switch_win = 0
stay_win = 0

win_ratios = []

for _ in range(100000):
    # Initialize problem
    x = [1, 0, 0]
    state = [i for i in range(len(x))]
    shuffle(x)

    # Choose door
    guess = randrange(0, 2)
    state.remove(guess)

    # Game show host opens door with goat
    goat_indices = [idx for idx, value in enumerate(x) if
                    value == 0 and idx != guess]
    open_door_idx = randrange(0, len(goat_indices))
    open_door = goat_indices[open_door_idx]
    state.remove(open_door)
```

Problem 6 continued on next page. . . 

```

    if x[guess] == 1:
        stay_win += 1
    else:
        switch_win += 1

    win_ratios.append((switch_win/(switch_win + stay_win),
                       stay_win/(switch_win + stay_win)))

print("switch:", switch_win/(switch_win + stay_win))
print("stay:", stay_win/(switch_win + stay_win))

switch_points = [(x, y[0]) for x, y in enumerate(win_ratios)]
stay_points = [(x, y[1]) for x, y in enumerate(win_ratios)]
x, = plt.plot(*zip(*switch_points))
y, = plt.plot(*zip(*stay_points))
plt.legend([x, y], ["Switch win %", "Stay win %"])
plt.xlabel('Games played')
plt.ylabel('Probability of winning')
plt.show()
```