Worksheet on “Decision Theory (incl. Bayes Classifiers)”

CS5691 PRML Jul–Nov 2025

August 13, 2025 (Version 2)

1. Consider a continuous random variable X and a discrete random variable Y. Let

- $P_Y(Y = 1) = 0.5$ and $P_Y(Y = -1) = 0.5$, and
- $(X|Y = 1) \sim \text{Unif}(-1, 2)$ and $(X|Y = -1) \sim \text{Unif}(-2, 1)$.

a. What is the marginal distribution of X? Specifically, plot the pdf of X denoted $f_X(x)$.

b. Write down the pdf $f_X(x)$ of X.

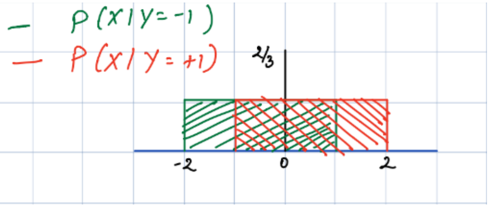
c. Write down and plot the posterior $Y|X$.

d. What is the optimal classifier for predicting Y from X, given the above assumptions?

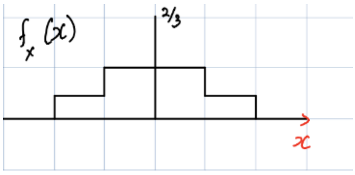
Solution:

a. Solution

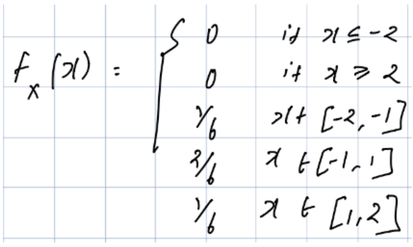
Class conditionals $f_{X|Y}(x)$:



Marginal $f_X(x)$:



b. See Figure.



$$f_X(x) = \begin{cases} 0 & \text{if } x \le -2 \\ 0 & \text{if } x \geq 2 \\ 1/6 & x \in [-2,-1] \\ 2/6 & x \in [-1,1] \\ 1/6 & x \in [1,2] \end{cases}$$

[Source: Parts a,b from [HG]Notes (by Harish Guruprasad)]

c. We use Bayes' theorem to derive the posterior as follows: $P(Y = 1 \mid X = x) = \frac{p(Y=1)f_{X|Y=1}(x)}{f_X(x)} = \frac{0.5\text{Unif}(x \mid [-1,2])}{f_X(x)}$.

$$P(Y = 1 \mid X = x) = \begin{cases} 0/0 & \text{if } x < -2 \text{ or } x > 2 \text{ (not in the support of } X) \\ 0 & \text{if } -2 \leq x < -1 \\ 0.5 & \text{if } -1 \leq x \leq 1 \\ 1 & \text{if } 1 < x \leq 2 \end{cases}$$

d. Using the posterior computed in the previous part, we can obtain the optimal classifier $h^*(x)$ with respect to the standard 0-1 loss function as:

$$h^*(x) = \begin{cases} 1 & \text{if } Pr(Y = 1|X = x) > Pr(Y = -1|X = x)\text{, and} \\ -1 & \text{otherwise} \end{cases}$$

Substituting the posterior in the above equation and choosing arbitrarily (in the case of tie where $Pr(Y = 1|X = x) == Pr(Y = -1|X = x)$ or in the case of invalid support), we get:

$$h^*(x) = \begin{cases} -1 & \text{if } x < -2 \text{ or } x > 2 \text{ (not in the support of } X) \\ -1 & \text{if } -2 \leq x < -1 \\ 1 & \text{if } -1 \leq x \leq 1 \text{ (tie)} \\ 1 & \text{if } 1 < x \leq 2 \end{cases}$$

2. Derive the Bayes classifier for binary classification ($Y = \pm 1$) under the below assumptions:

$$P(Y = 1) = 0.7 \text{ and } P(Y = -1) = 0.3$$
$$X|Y = 1 \sim Unif(-1, 3)$$
$$X|Y = -1 \sim Unif(-2, 0)$$

Solution:

$$\eta(x) = P(y=1|X=x) = \frac{f_{X|Y}(x|1) \cdot 0.7}{0.7 f_{X|Y}(x|1) + 0.3 f_{X|Y}(x|-1)}$$

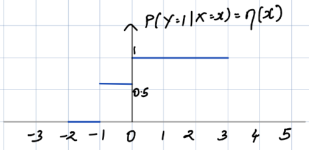
Case 1 : $x < -2 \Rightarrow \eta(x) = \frac{0}{0}$ = Doesn't matter

Case 2 : $x \in [-2,-1) \Rightarrow \eta(x) = \frac{0}{0+\frac{1}{2}(0.3)} = 0$

Case 3 : $x \in [-1,0) \Rightarrow \eta(x) = \frac{1/4 \cdot (0.7)}{1/4 (0.7) + 1/2 (0.3)} =$

Case 4 : $x \in [0,3) \Rightarrow \eta(x) = \frac{1/4 (0.7)}{1/4 (0.7)} = 1$

Case 5 : $x \geqslant 3 \Rightarrow \eta(x) = \frac{0}{0}$ = Doesn't matter



$$h^*(x) = \begin{cases} -1 & x \in [-2,-1) \\ +1 & x \in [-1,3] \\ \text{Don't care} & \text{o.w} \end{cases}$$

[Source: From [HG]Notes]

3. [Link theory to practice] For a binary classifer h, let $L = \begin{bmatrix} p & q \\ r & s \end{bmatrix}$ be the loss matrix; and $C_{\text{train}} = \begin{bmatrix} 100 & 10 \\ 20 & 120 \end{bmatrix}$, and $C_{\text{test}} = \begin{bmatrix} 90 & 45 \\ 30 & 85 \end{bmatrix}$ be the confusion matrix when h is applied on the training and test data respectively. All three matrices have ground-truth classes t along the rows and predictions h along the columns in the same order for the two classes. Express your estimate of the risk (expected loss) of h in terms of p to s above.

Solution: For any random variable (rv) Y with pdf $f_Y(.)$, we can estimate the expectation of a function $g(.)$ of Y as the average of this function evaluated at N iid samples from Y. That is,

$$\begin{aligned} \mathbb{E}_Y[g(Y)] &= \int_y g(y)\ f_Y(y)dy \\ &\approx \frac{1}{N}\sum_{i=1}^{N} g(y_i), \text{where each } y_i \sim_{iid} f_Y(.). \end{aligned}$$

This technique can also be used to estimate the expectation of the loss rv $L_{t,h(x)}$, which in turn is a function of the rvs x, t (feature vector x and target label t with true joint distribution $f_{x,t}(.,.)$). Thus, under the assumption that the test dataset is iid sampled from $f_{x,t}(.,.)$,

$$\begin{aligned} \mathbb{E}_{x,t}[L(x,t)] &= \int_x \sum_{t\in\{C_1,C_2\}} L_{t,h(x)} f_{\mathbf{x},\mathbf{t}}(x,t)dx \\ &\approx \frac{1}{|\text{test-dataset}|} \sum_{(x_i,t_i)\in\text{test-dataset}} L_{t_i,h(x_i)} \\ &= \frac{1}{250}(90p + 45q + 30r + 85s). \end{aligned}$$

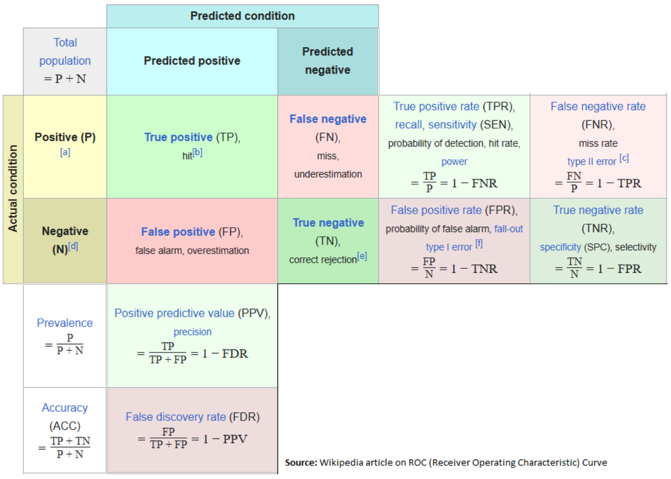
We used the test-dataset here instead of the training dataset, because all the rvs $\{h(x_i)\}_{i\in D}$ and in turn $\{L_{t_i,h(x_i)}\}_{i\in D}$ are independent of each other if D is the test-dataset (with independence of the L samples resulting in more reliable estimate of its expectation).

The training dataset is also iid sampled from $f_{x,t}(.,.)$ like the test dataset, but the rvs $\{h(x_i)\}_{i\in D}$ are **not** independent of each other, because the $h(.)$ classifier is learnt using information from all the training datapoints. Having said that and despite this lack of independence of rvs, the risk calculated from the training data (also known as empirical risk) is also useful, but for a different purpose (specifically for training purposes, where empirical risk minimization (ERM) approach is often used to choose the classifier with the lowest risk in training data).

To summarize, training-dataset-based risk estimate is used to learn the optimal classifier during the training phase, and unseen-test-dataset-based risk estimate is a reliable estimate of the risk of this optimal classifier (or any other classifier of interest).

4. [Link practice to theory] Besides expected loss, many other performance metrics can help evaluate the quality of a binary classifier h in practice (see figure below beside the confusion matrix).

Consider these performance/evaluation metrics of h: **Precision, Recall/Sensitivity, and Specificity**. The formula given in the figure for these metrics is actually a test-dataset-based estimate of a probability (defined over the probability space $(\mathbf{x}, t)$, where $\mathbf{x}$ is the input and $t \in \{-1, +1\}$ is the binary target). Write down this probability, i.e., **express each of these three metrics for a classifier $h(\mathbf{x})$ as a probability** over the joint probability space of $(\mathbf{x}, t)$.

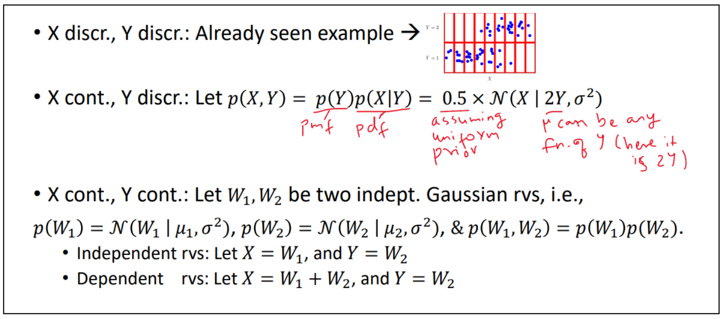


		Predicted condition			
	Total population $= \mathrm{P} + \mathrm{N}$	Predicted positive	Predicted negative		
Actual condition	Positive (P) [a]	True positive (TP), hit[b]	False negative (FN), miss, underestimation	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{\mathrm{TP}}{\mathrm{P}} = 1 - \mathrm{FNR}$	False negative rate (FNR), miss rate type II error [c] $= \frac{\mathrm{FN}}{\mathrm{P}} = 1 - \mathrm{TPR}$
	Negative (N)[d]	False positive (FP), false alarm, overestimation	True negative (TN), correct rejection[e]	False positive rate (FPR), probability of false alarm, fall-out type I error [f] $= \frac{\mathrm{FP}}{\mathrm{N}} = 1 - \mathrm{TNR}$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{\mathrm{TN}}{\mathrm{N}} = 1 - \mathrm{FPR}$
	Prevalence $= \frac{\mathrm{P}}{\mathrm{P} + \mathrm{N}}$	Positive predictive value (PPV), precision $= \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} = 1 - \mathrm{FDR}$			
	Accuracy (ACC) $= \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{P} + \mathrm{N}}$	False discovery rate (FDR) $= \frac{\mathrm{FP}}{\mathrm{TP} + \mathrm{FP}} = 1 - \mathrm{PPV}$			

Source: Wikipedia article on ROC (Receiver Operating Characteristic) Curve

Solution: Probabilistic view of metrics:
Precision $= Pr_{\mathrm{x},t}(t = 1 \mid h(\mathrm{x}) = 1)$
Recall or Sensitivity $= Pr_{\mathrm{x},t}(h(\mathrm{x}) = 1 \mid t = 1)$
Specificity $= Pr_{\mathrm{x},t}(h(\mathrm{x}) = -1 \mid t = -1)$

5. Consider the four examples of two jointly distributed rvs (X, Y) from Slide 19 of "M0a. Background on Probabillity", a screenshot of which is shown below. For each of these examples, write down the optimal (Bayes) classifier for predicting Y given X (in case of discrete Y) and optimal regressor for predicting Y given X (in case of continuous Y). Assume that standard loss functions (0-1 loss function for classification and squared loss function for regression) need to be optimized.



- X discr., Y discr.: Already seen example →
- X cont., Y discr.: Let $p(X,Y) = \underbrace{p(Y)}_{\text{pmf}}\underbrace{p(X|Y)}_{\text{pdf}} = \underbrace{0.5}_{\text{assuming uniform prior}} \times \mathcal{N}(X \mid \underbrace{2Y}_{\mu \text{ can be any fn. of } Y \text{ (here it is } 2Y)}, \sigma^2)$
- X cont., Y cont.: Let W_1, W_2 be two indept. Gaussian rvs, i.e.,

$p(W_1) = \mathcal{N}(W_1 \mid \mu_1, \sigma^2)$, $p(W_2) = \mathcal{N}(W_2 \mid \mu_2, \sigma^2)$, & $p(W_1, W_2) = p(W_1)p(W_2)$.
  - Independent rvs: Let $X = W_1$, and $Y = W_2$
  - Dependent rvs: Let $X = W_1 + W_2$, and $Y = W_2$

Solution:

a. Given x, the Bayes classifier $h^*(x)$ is set to the value j of Y that has the highest number of dots for that value x of X. That is,

$$h^*(x) = \arg\max_j Pr(Y = j | X = x).$$

This yields:

$$h^*(x) = \begin{cases} 1 \text{ if } x \in \{1, \ldots, 5\} \\ 2 \text{ if } x \in \{6, \ldots, 9\} \end{cases}$$

b. The class prior $Pr(Y)$ is uniform (0.5), so $p(X,Y) = p(Y)p(X|Y) = 0.5\,p(X|Y)$. Now, $(X|Y = 1)$ and $(X|Y = 2)$ are both Gaussian distributed with same variance and mean 2 and 4 respectively. Due to their same variance, the pdfs of these two rvs will meet midway at $x = 3$, with $0.5\,p(X|Y = 1)$ dominating $0.5\,p(X|Y = 2)$ for x values lower than 3 and vice versa otherwise. This yields the following Bayes classifier:

$$h^*(x) = \begin{cases} 1 \text{ if } x < 3 \\ 2 \text{ if } x \geq 3 \end{cases}$$

c. **Independent rvs**: We know the optimal regressor is $f^*(x) = E[Y|X = x]$. Therefore,

$$\begin{aligned} f^*(x) &= E[Y|X = x] \\ &= E[Y] \quad \text{(because } Y \text{ is indep. of } X\text{)} \\ &= E[W_2] \\ &= \mu_2 \end{aligned}$$

d. **Dependent rvs**: We know the optimal regressor $f^*(x)$ is $E[Y|X = x]$. We can derive this conditional expectation from known results of a multi-variate Gaussian (MVG) random vector (which we will see later in the class, and which you can also find in Slide 58 of M0a slide deck or CMB book Chapter 2 section on "Partitioned Gaussians"). Applying this result on partitioned Gaussians to the MVG distributed random vector (X, Y), it follows that:

$$\begin{aligned} (Y|X = x) &\sim \mathcal{N}\left(\frac{(x-\mu_1)+\mu_2}{2}, \frac{\sigma^2}{2}\right) \text{ . So,} \\ f^*(x) &= E[Y|X = x] = \frac{(x-\mu_1)+\mu_2}{2}. \end{aligned}$$

Alternatively, we can derive the above conditional distribution, and hence conditional expectation, directly from first principles (i.e., without resorting to known MVG results on partitioned Gaussians) as shown below. Let $p(.)$ denote the pdf of the corresponding rv, and $\mathcal{N}(z|\mu, \sigma^2)$ denote the pdf of a Gaussian rv Z with mean μ and variance σ^2 evaluated at $Z = z$.

$$\begin{aligned} p(Y = y \mid X = x) &= p(W_2 = y \mid W_1 + W_2 = x) \quad \text{(by defn. of } X, Y\text{)} \\ &= \frac{p(W_2 = y, W_1 + W_2 = x)}{p(W_1 + W_2 = x)} \\ &= \frac{p(W_2 = y, W_1 + y = x)}{p(W_1 + W_2 = x)} \\ &= \frac{p(W_2 = y)p(W_1 = x - y)}{p(W_1 + W_2 = x)} \quad \text{(by indep. of } W_1, W_2\text{)} \\ &= \frac{p(W_2 = y)\,p(W_1 = x - y)}{\mathcal{N}(x \,|\, \mu_1 + \mu_2,\, 2\sigma^2)} \quad \text{(as sum of two indept. Gaussian rvs is also Gaussian)} \\ &= \frac{\mathcal{N}(y \,|\, \mu_2,\, \sigma^2)\,\mathcal{N}(x - y \,|\, \mu_1,\, \sigma^2)}{\mathcal{N}(x \,|\, \mu_1 + \mu_2,\, 2\sigma^2)} \\ &= \frac{\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(y-\mu_2)^2}{2\sigma^2}\right\} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-y-\mu_1)^2}{2\sigma^2}\right\}}{\frac{1}{\sqrt{2\sigma^2}\sqrt{2\pi}} \exp\left\{-\frac{(x-(\mu_1+\mu_2))^2}{2(2\sigma^2)}\right\}} \\ &= \frac{\sqrt{2}}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(y-\mu_2)^2 + (x-y-\mu_1)^2 - \frac{(x-(\mu_1+\mu_2))^2}{2}}{2\sigma^2}\right\} \\ &= \frac{1}{\sqrt{\sigma^2/2}\sqrt{2\pi}} \exp\left\{-\frac{(y-\mu_2)^2 + (y-(x-\mu_1))^2 - \frac{((x-\mu_1)-\mu_2)^2}{2}}{2\sigma^2}\right\} \\ &= \frac{1}{\sqrt{\sigma^2/2}\sqrt{2\pi}} \exp\left\{-\frac{(y - \frac{(x-\mu_1)+\mu_2}{2})^2}{2(\sigma^2/2)}\right\} \quad \text{(upon algebraic manipulations,} \\ &\qquad\qquad \text{specifically completing the squares)} \\ &= \mathcal{N}\left(y \;\middle|\; \frac{(x-\mu_1)+\mu_2}{2},\, \frac{\sigma^2}{2}\right). \end{aligned}$$

■