

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?

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

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

3. [LINK THEORY TO PRACTICE] For a binary classifier 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.


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 = $P + N$ | 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 $= \frac{TP}{P} = 1 - \text{FNR}$ | False negative rate (FNR), miss rate type II error [c] $= \frac{FN}{P} = 1 - \text{TPR}$ |
| | Negative (N) [d] | False positive (FP), false alarm, overestimation | True negative (TN), correct rejection [e] | False positive rate (FPR), probability of false alarm, tail-out type I error [f] $= \frac{FP}{N} = 1 - \text{TNR}$ | True negative rate (TNR), specificity (SPC), selectivity $= \frac{TN}{N} = 1 - \text{FPR}$ |
| Prevalence $= \frac{P}{P + N}$ | | Positive predictive value (PPV), precision $= \frac{TP}{TP + FP} = 1 - \text{FDR}$ | | | |
| Accuracy (ACC) $= \frac{TP + TN}{P + N}$ | | False discovery rate (FDR) $= \frac{FP}{TP + FP} = 1 - \text{PPV}$ | | | |

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

5. Consider the four examples of two jointly distributed rvs (X, Y) from Slide 19 of “M0a. Background on Probability”, 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 $p(X, Y) = \underbrace{p(Y)}_{\text{pmf}} \underbrace{p(X|Y)}_{\text{pdf}} = 0.5 \times \mathcal{N}(X | 2Y, \sigma^2)$
assuming uniform prior *p can be any fn. of Y (here it is 2Y)*
- X cont., Y cont.: Let W_1, W_2 be two indept. Gaussian rvs, i.e.,
 $p(W_1) = \mathcal{N}(W_1 | \mu_1, \sigma^2)$, $p(W_2) = \mathcal{N}(W_2 | \mu_2, \sigma^2)$, & $p(W_1, W_2) = p(W_1)p(W_2)$.
 - Independent rvs: Let $X = W_1$, and $Y = W_2$
 - Dependent rvs: Let $X = W_1 + W_2$, and $Y = W_2$