Risks and Bayes Optimal Prediction

Song Liu (song.liu@bristol.ac.uk)

Office Hour: Thursday 3-4pm

Office: GA 18

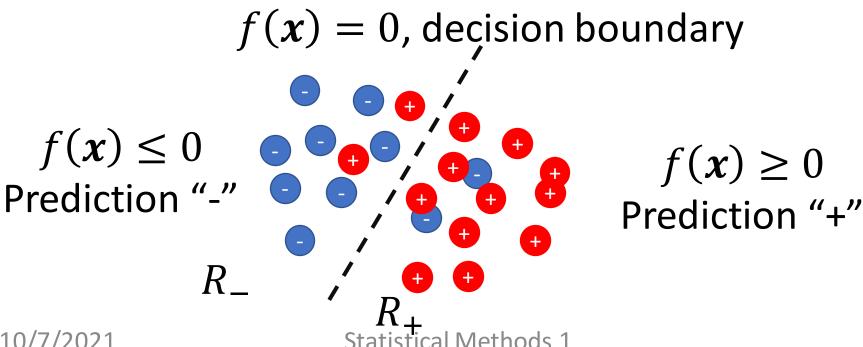


Binary Classification

- Sometimes, we need to make discrete decisions
 - In contrast to regression which only predicts a continuous value.
 - e.g., given X-ray image of a person, we decide whether this person is a sick or not.
- Output: $y \in \{+1, -1\}$, class label.
 - A binary decision of class, e.g., "normal" or "patient"
- Input: $x \in R^d$
 - The input, such as an X-ray image of a person.
- Task: Given x make a prediction y
- We want to make as little mistakes as possible.

Binary Classification

- Rather than fit a function like we did in regression, in binary classification, we look for a decision boundary, which separates space of x into two areas R_+ and R_- .
- A decision boundary is defined by a function f(x)



False Positive and False Negative

- What is the best f(x) given a dataset D?
- To answer this question, we need to know what are the mistakes we can make in a binary classification.
 - False positive (FP): an x should have been labelled "-1", but is labelled "+1".
 - False negative (FN): an x should have been labelled "+1", but is labelled "-1".
 - Similarly, we can define True Positive (TP) and True Negative (TN).

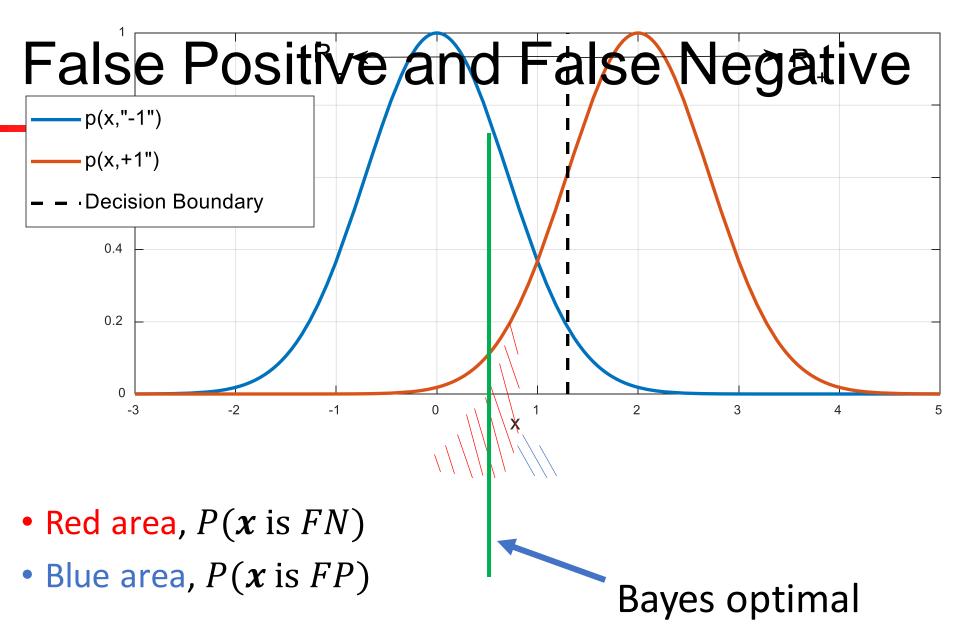
False Positive and False Negative

- Let us look at this problem from a probabilistic perspective:
- Probability density of "+" data: p(x|y = " + 1")
- Probability density of "-" data: p(x|y = "-1")
- Probability of class itself, p(y = +1) and p(y = -1).
- What is the probability of making mistakes given areas R_+ and R_- create by a decision function f(x)?
- P(x is FP or FN|f)= $\int_{R_{+}} p(x, y = "-1") dx + \int_{R_{-}} p(x, y = "+1") dx$
- Prove: $P(FP \ or \ FN|f)$ is minimized when
- f(x) = p(x, y = +1) p(x, y = -1).

Bayes Optimal Classifier

- f(x) = p(x, y = +1) p(x, y = -1)
- In literatures, this f is referred as Bayes optimal classifier.

- However, this only serves as an idealized optimal classifier.
- In reality, we do not have access to p(x, y) but only data points $D = \{(x_i, y_i)\}_{i=1}^n$.
 - Infer joint distribution p(x, y) from data is usually very hard.
 - We will see two different strategies later which can be used to ease the difficulty.



Risks in Decision Making

- Making wrong decisions may have different loss.
- We might weight FP and FN differently.

- For example, diagnosing a patient as healthy (FN) is certainly riskier than diagnosing a healthy person as a patient (FN).
 - The patient may miss his/her treatment.
 - Treating a healthy person is usually less dangerous.

Patient Treatment Loss Matrix

 Imagine we can quantify the cost of decision making using a loss matrix.

•
$$L = \begin{bmatrix} patient & normal \\ patient \\ normal \end{bmatrix}$$

- It says, if we label a patient as a normal person, the cost is 1000 times as labelling a normal person as patient.
 - We pay no price for correct labelling.
- Giving this loss matrix, how to make a good cost-sensitive decision?

Risk Minimization

- To make a good decision, we need to minimize the expected loss of making a wrong decision.
- Suppose output is $y \in \{\text{normal}, \text{patient}\}$, and input is x
- Given x, a decision is $y_0 \in \{\text{normal, patient}\}\$
- Then the optimal decision is given by

$$\operatorname{argmin}_{y_0} \mathbb{E}_{p(y|x)}[L(y,y_0) | x]$$

- Where L is a function whose value is determined by L.
 - e.g. $L(y = \text{normal}, y_0 = \text{patient}) = 1$

Risk Minimization

- As y is a discrete variable, we can write down $\mathbb{E}_{p(y|x)}[L(y,y_0) \mid x] = \sum_{y \in \{+1,-1\}} p(y|x) L(y,y_0)$
- The expectation is a **weighted sum of** $L(y, y_0)$, weighted by p(y|x).

- Problem: we cannot compute this weighted sum, as
- We have no idea what is p(y|x).
- We can infer it from using a dataset D.

Inference of p(y|x)

- Replace p(y|x) with p(y|x,D)!
- The decision is now given by
- $\operatorname{argmin}_{y_0} \mathbb{E}_{p(y|x,D)}[L(y,y_0)|x]$
- Problem: How to get p(y|x,D)?
 - MLE, MAP, Full Probabilistic Approach blablabla...

Calculate p(y|x,D)

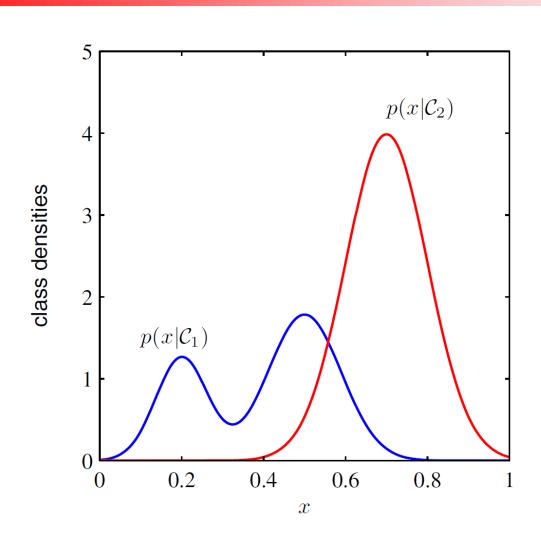
- In classification tasks, there are two schools of thoughts on how to obtain p(y|x,D), both have pros and cons.
- A straightforward approach.
 - Infer p(y|x, D) directly.
- An indirect approach: $p(y|x,D) \propto p(x|y,D)p(y)$.
 - Infer p(x|y, D) using D.
 - p(y = +1) and p(y = -1) is just the proportion of pos/neg samples.
- The inference of p(y|x,D) or p(x|y,D) can be done using MLE, MAP or full probabilistic methods, we will touch this later.

Discriminative vs. Generative

- Straightforward approach models p(y|x) with p(y|x; w).
 - This is called **discriminative** approach.
 - p(y|x) only tells the difference between pos/neg.
 - It does not allow us to simulate new x given a class y.

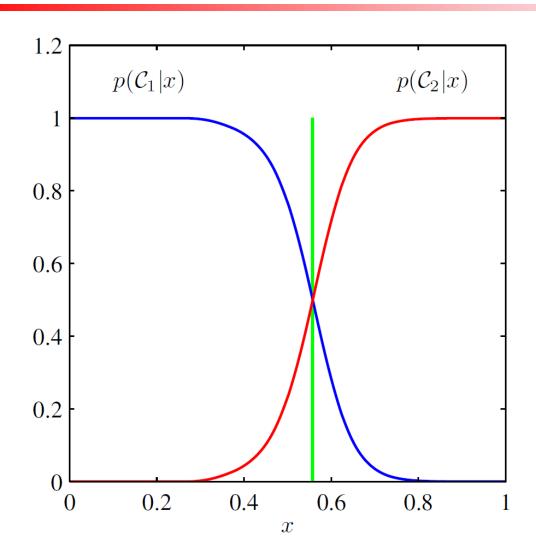
- Indirect approach models p(x|y) with p(x|y; w) instead.
 - This is called **generative** approach.
 - p(x|y) can "generate" new input x given an output y.
 - Learning a p(x|y) with a high dim. x can still be difficult.

Class Densities: p(x|y)



• PRML, Figure 1.27

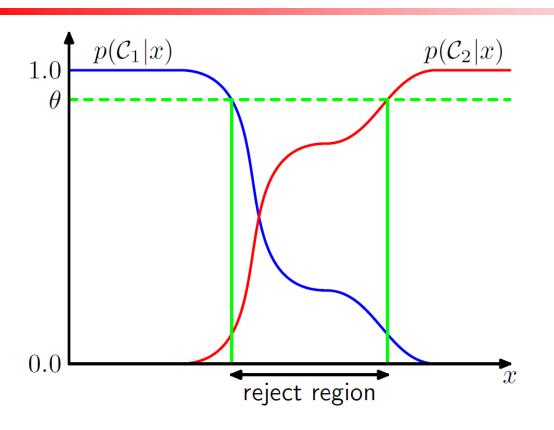
Class-Posterior Probability: p(y|x)



• PRML, Figure 1.27

17

Rejection Option



PRML Figure 1.26

We can reject decision making when we find $\max\{p(y=+1|x), p(y=-1|x)\}$ is lower than a threshold.

What about Regression?

- Output value of regression is a continuous variable.
 - We cannot have a loss matrix anymore.

- We can use the loss function, such as squared-loss
- $L(y, y_0) = (y y_0)^2$
- Again, we minimize the expected loss:
- \hat{y} : = $\underset{y_0}{\operatorname{argmin}} \mathbb{E}_{p(y|x)}[L(y,y_0)|x]$ Prove the optimal prediction using squared-loss is
- $\hat{y} := \mathbb{E}_{p(y|x)}[y].$

What about Regression?

•
$$\hat{y} \coloneqq \mathbb{E}_{p(y|\mathbf{x})}[y]$$

- We do not have p(y|x), but we can infer p(y|x,D).
- $\hat{y} \approx \mathbb{E}_{p(y|\mathbf{x},D)}[y]$
- The optimal prediction with respect to squared-risk function is calculating the mean of the inferred p(y|x,D).

Absolute Value Risk Function

- We can use other loss functions, e.g. absolute val. function
 - $L(y, y_0) \coloneqq |y y_0|$
- Prove the optimal prediction using absolute value loss is:
- argmin_{y_0} $\mathbb{E}_{p(y|x)}[|y-y_0|]$ is the Median of p(y|x).
- Median m is defined as a real value such that

•
$$\int_{-\infty}^{m} p(y|x) dy = \int_{m}^{+\infty} p(y|x) dy = \frac{1}{2}$$

• Or the "50% percentile".

Computing Lab (1)

- Generate data $y_i = \exp(1.5x_i 1) + \epsilon_i$, $\epsilon_i \sim N(0, .64)$.
 - $i = 1 \dots 200$

- Modify your last week's implementation of least squares to calculate the regularized least squares solution: $w_{\rm LS-R}$.
- Tuning regularization constant λ and measure the CV error.

• Can you find a λ such that CV error is minimized?

Computing Lab (2)

- Using the same dataset,
- Calculate the predictive probability distribution using the "marginalization trick":
 - $p(\hat{y}|x,D)$
- Plot $\mathbb{E}_{p(\widehat{y}|x,D)}[\widehat{y}|x]$ on your dataset, as a function of x.
- Plot "the tube",
- $\mathbb{E}_{p(\hat{y}|\boldsymbol{x},D)}[\hat{y}|\boldsymbol{x}] \pm \sqrt{\operatorname{var}_{p(\hat{y}|\boldsymbol{x},D)}[\hat{y}|\boldsymbol{x}]}$
- How much data does the tube cover (in terms of percentage)?