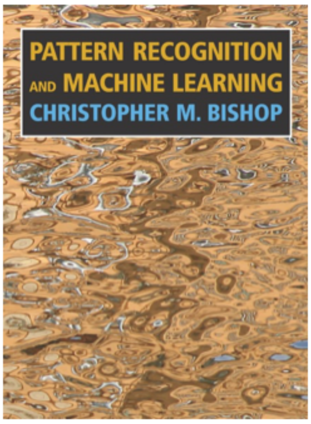
Risks and Bayes Optimal Prediction

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

Office: GA 18

Reference



Today's class *roughly* follows Chapter 1.

Pattern Recognition and Machine Learning

Christopher Bishop, 2006

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**: $\boldsymbol{x} \in R^d$
  - The input, such as an X-ray image of a person.

- **Task**: Given $\boldsymbol{x}$ make a prediction y
- We want to make as little mistakes as possible.

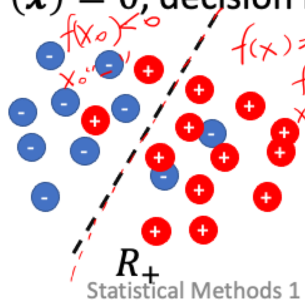
Binary Classification

$f(x)$ 拟合

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

$f(\boldsymbol{x}) = 0$, decision boundary

$f(\boldsymbol{x}) \leq 0$
Prediction "-"



$f(\boldsymbol{x}) \geq 0$
Prediction "+"

R_-

R_+

False Positive and False Negative

- What is the best $f(\boldsymbol{x})$ 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 $\boldsymbol{x}$ should have been labelled "-1", but is labelled "+1".
  - False negative (FN): an $\boldsymbol{x}$ 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: $p(\boldsymbol{x}|y = "+1")$
- Probability density of "-" data: $p(\boldsymbol{x}|y = "-1")$
- 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 $f(x)$?
- $P(\boldsymbol{x} \text{ is } FP \text{ or } FN|f)$

FP FN

$$= \int_{R_+} p(\boldsymbol{x}, y = "-1")\mathrm{d}\boldsymbol{x} + \int_{R_-} p(\boldsymbol{x}, y = "+1")\mathrm{d}\boldsymbol{x}$$

- Prove: $P(FP \text{ or } FN|f)$ is minimized when
- $f(\boldsymbol{x}) = p(\boldsymbol{x}, y = +1) - p(\boldsymbol{x}, y = -1)$.

10/16/20 Statistical Methods 1 6

$p(\boldsymbol{x}, y = -1)$ is the likelihood of x being negative
$p(\boldsymbol{x}, y = 1)$ is the likelihood of x being positive

Therefore, integrating them over the positive and negative region gives you the probability of making mistakes (FP, FN)

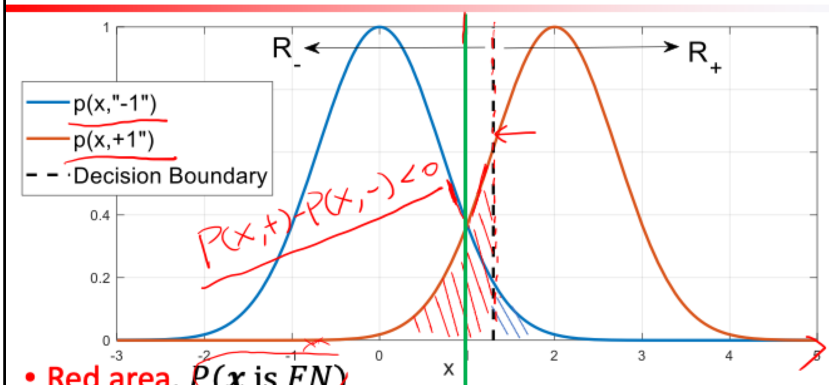
Bayes Optimal Classifier

- $f(\boldsymbol{x}) = p(\boldsymbol{x}, y = +1) - p(\boldsymbol{x}, y = -1)$
- 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 $p(\boldsymbol{x}, y)$ but only data points $D = \{(x_i, y_i)\}_{i=1}^{n}$.
  - Infer joint distribution $p(\boldsymbol{x}, y)$ from data is usually very hard.
  - We will see two different strategies later which can be used to ease the difficulty.

Joint probability completely characterize the data generating source. Using limited data points to infer such a strong result is usually hard.

False Positive and False Negative



- Red area, $P(\boldsymbol{x} \text{ is } FN)$
- Blue area, $P(\boldsymbol{x} \text{ is } FP)$

Bayes optimal

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 (FP).
  - 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**.
- $\boldsymbol{L} = \begin{array}{c c} & \begin{array}{cc} \text{patient} & \text{normal} \end{array} \\ \begin{array}{c} \text{patient} \\ \text{normal} \end{array} & \begin{bmatrix} 0 & 1000 \\ 1 & 0 \end{bmatrix} \end{array}$
- 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?

10/16/20 Statistical Methods 1 10

We only have four different scenarios, TP, TN, FP, FN, which corresponding to the four different entries of the loss matrix.

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 $\boldsymbol{x}$
- Given $\boldsymbol{x}$, a decision is $y_0 \in \{\text{normal}, \text{patient}\}$
- Then the optimal decision is given by

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

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

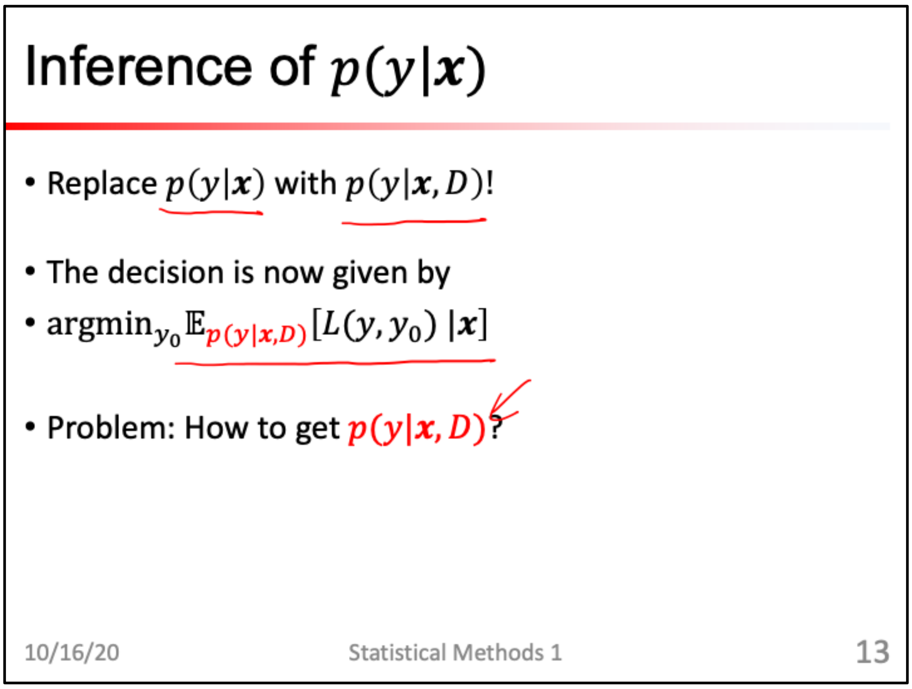
The dataset is random, so we do not really care about an individual decision is right or wrong. Instead, we care the expected loss.

Risk Minimization

- As y is a discrete variable, we can write down $\mathbb{E}_{p(y|\boldsymbol{x})}[L(y, y_0)\ |\boldsymbol{x}] = \sum_{y\in\{+1,-1\}} p(y|\boldsymbol{x})L(y, y_0)$
- The expectation is a **weighted sum of** $L(y, y_0)$**,** weighted by $p(y|\boldsymbol{x})$.
- **Problem**: we cannot compute this weighted sum, as
- We have no idea what is $p(y|\boldsymbol{x})$.
- We can infer it from using a dataset D.

10/16/20 Statistical Methods 1 12

inferring $p(y|\boldsymbol{x})$ is usually much easier than inferring $p(y, \boldsymbol{x})$ from the dataset, the full joint probability.



We do not have $p(y|x)$, we have the next best thing, $p(y|x, D)$

Calculate $p(y|\boldsymbol{x}, D)$

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

10/16/20 Statistical Methods 1 14

For classification tasks only. The reasoning here in general does not apply to regression tasks.

Discriminative vs. Generative

- Straightforward approach models $p(y|\boldsymbol{x})$ with $p(y|\boldsymbol{x};\boldsymbol{w})$. 0.2
  - This is called **discriminative** approach.
  - $p(y|\boldsymbol{x})$ only tells the difference between pos/neg.
  - It does not allow us to simulate new $\boldsymbol{x}$ given a class y.

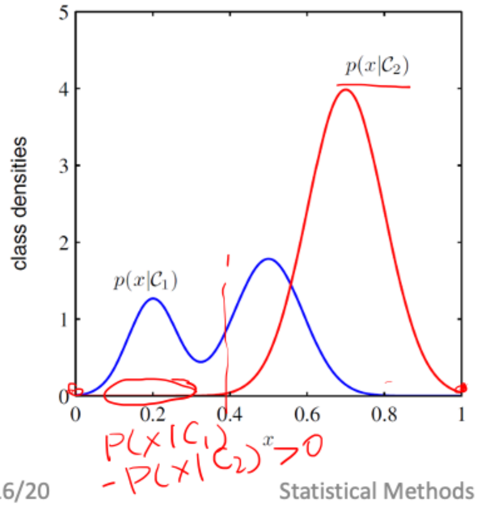
- Indirect approach models $p(\boldsymbol{x}|y)$ with $p(\boldsymbol{x}|y;\boldsymbol{w})$ instead.
  - This is called **generative** approach.
  - $p(\boldsymbol{x}|y)$ can "generate" new input $\boldsymbol{x}$ given an output y.
  - Learning a $p(\boldsymbol{x}|y)$ with a high dim. $\boldsymbol{x}$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

The probability space of p(y|x) is only binary, but the probability space of p(x|y) is a much bigger space.

If your tasks is only to classify data points, making discrete decisions, usually the discriminative approach is your best bet.

Learning p(x|y) requires you to model and infer a high dimensional distribution on x, which usually suffers from the curse of dimensionality.

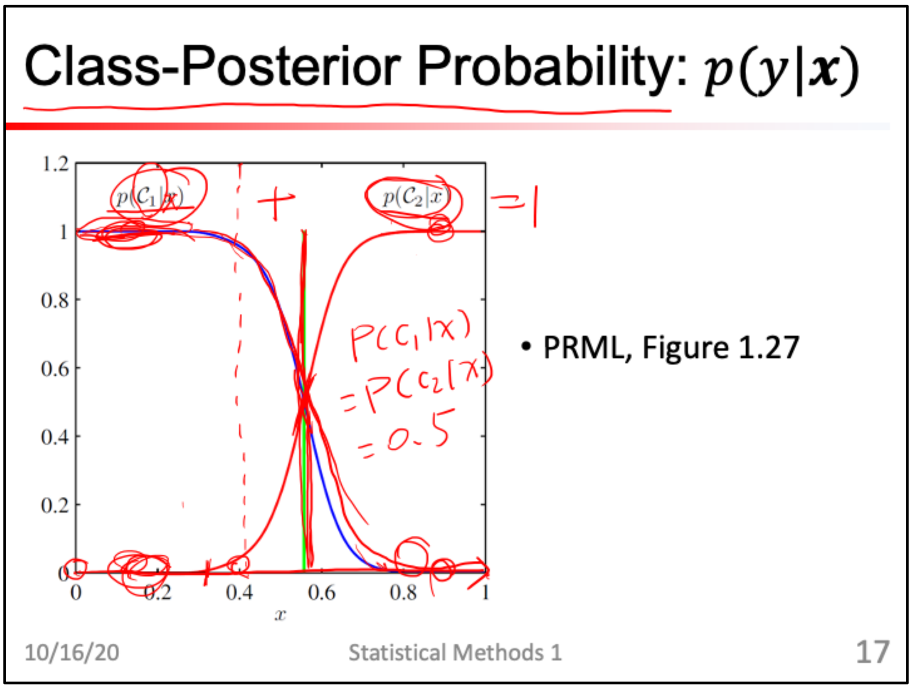
Class Densities: $p(\boldsymbol{x}|y)$



- PRML, Figure 1.27

10/16/20 Statistical Methods 1 16

It tells you how x is distributed at the interval [0,1]

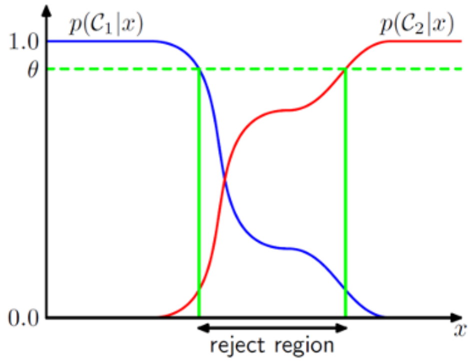


They look very differently from class densities, as they are probabilities of binary variables.

It tells you how likely x is in blue/red class give different x on the horizontal axis.

You can see the class posterior probability looks a lot simpler, cleaner!! This is why, if your task is only classification, discriminative approach is your best bet.

Rejection Option



PRML **Figure 1.26**

We can reject decision making when we find $\max\{p(y=+1|\boldsymbol{x}), p(y=-1|\boldsymbol{x})\}$ 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:
- $\hat{y} := \mathrm{argmin}_{y_0} \mathbb{E}_{p(y|\boldsymbol{x})}[L(y, y_0)\ |\boldsymbol{x}]$
  $= \mathrm{argmin}_{y_0} \mathbb{E}_{p(y|\boldsymbol{x})}[(y - y_0)^2 |\boldsymbol{x}]$
- Prove: $\hat{y} \coloneqq \mathbb{E}_{p(y|\boldsymbol{x})}[y]$.

What about Regression?

- $\hat{y} := \mathbb{E}_{p(y|\boldsymbol{x})}[y]$

- We do not have $p(y|\boldsymbol{x})$, but we can have $p(y|\boldsymbol{x}, D)$.
- $\hat{y} \quad \approx \mathbb{E}_{p(y|\boldsymbol{x}, D)}[y]$
- $p(y|\boldsymbol{x}, D)$ can be inferred using MLE, MAP or Full Probabilistic approach, then the optimal prediction with respect to squared-risk function **corresponds to looking for the mean** of the inferred $p(y|\boldsymbol{x}, D)$.
  - When $p(y|\boldsymbol{x}, D)$ is inferred by MLE, least-squares give the optimal prediction.

Absolute Value Risk Function

- Prove:
- $\mathrm{argmin}_{y_0} \mathbb{E}_{p(y|x)}[\,|y - y_0|]$ is the Median of $p(\boldsymbol{y}|\boldsymbol{x})$.

- Median m is defined as a real value such that
- $\int_{-\infty}^{m} p(y|\boldsymbol{x})\,dy = \int_{m}^{+\infty} p(y|\boldsymbol{x})\,dy = \frac{1}{2}$

- 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: $\boldsymbol{w}_{\mathrm{LS-R}}$.
- 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":
  - $p(\hat{y}|\boldsymbol{x}, D)$

- Plot $\mathbb{E}_{p(\hat{y}|\boldsymbol{x}, D)}[\hat{y}|\boldsymbol{x}]$ on your dataset, as a function of $\boldsymbol{x}$.
- Plot "the tube",
- $\mathbb{E}_{p(\hat{y}|\boldsymbol{x}, D)}[\hat{y}|\boldsymbol{x}] \pm \sqrt{\mathrm{var}_{p(\hat{y}|\boldsymbol{x}, D)}[\hat{y}|\boldsymbol{x}]}$
- Assuming $\sigma, \sigma_w = 1$