Lecture 5

Basic Concepts in Bayesian Decision and Information Theory

Instructor: Shibo Li

shiboli@cs.fsu.edu

Department of Computer Science
Florida State University

Given $\mathbf{x}$, want to predict $\mathbf{t}$

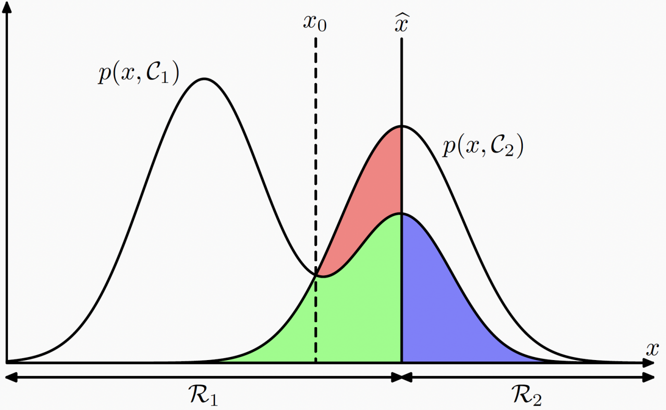
$\mathbf{t}$: Cancer, Stock price, Weather ...

- Inference step
  - Determine either p($\boldsymbol{t}$|$\boldsymbol{x}$) or p($\boldsymbol{x}$,$\boldsymbol{t}$) (from training data)
- Decision Step
  - For Given $\boldsymbol{x}$, determine optimal $\boldsymbol{t}$

Let us first consider the classification problem

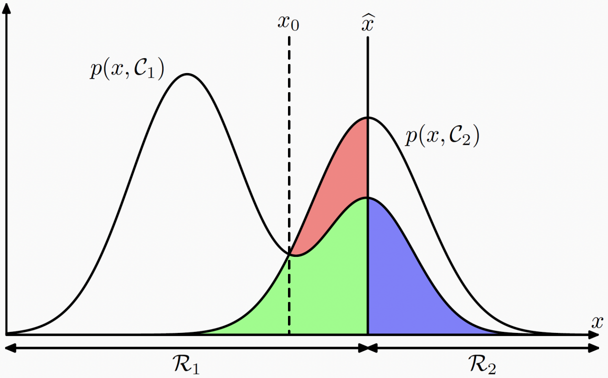
- $\boldsymbol{t} \in \{C_1, \ldots, C_K\}$
- Decision regions R_k: if $\boldsymbol{x}$ falls in , predict C_k
- Decision boundaries/surfaces: boundaries between different decision regions

Minimum Misclassification Rate



$$
\begin{aligned}
p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\
&= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2) \, \mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1) \, \mathrm{d}\mathbf{x}.
\end{aligned}
$$

Minimum Misclassification Rate



Question: where shall we set the decision boundary to minimize the misclassification rate? Why?

Minimum Misclassification Rate

- In general for *K* classes

$$\begin{aligned} p(\text{correct}) &= \sum_{k=1}^{K} p(\mathbf{x} \in \mathcal{R}_k, \mathcal{C}_k) \\ &= \sum_{k=1}^{K} \int_{\mathcal{R}_k} p(\mathbf{x}, \mathcal{C}_k)\, \mathrm{d}\mathbf{x} \end{aligned}$$

$$\boxed{p(\mathcal{C}_k|\mathbf{x})}\, p(\mathbf{x})$$

How to find regions that maximize the probability of correctness?

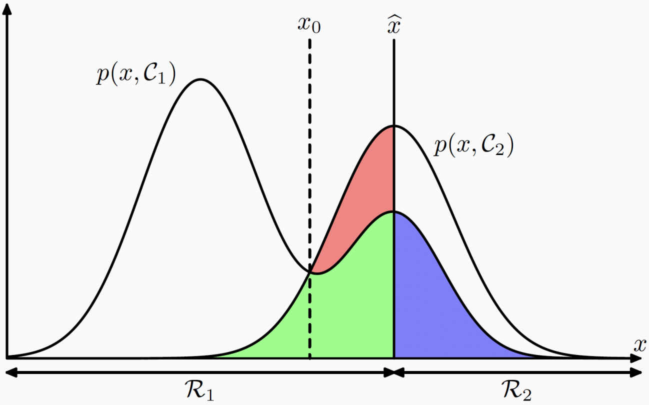
- In general for K classes

$$
\begin{aligned}
p(\text{correct}) &= \sum_{k=1}^{K} p(\mathbf{x} \in \mathcal{R}_k, \mathcal{C}_k) \\
&= \sum_{k=1}^{K} \int_{\mathcal{R}_k} p(\mathbf{x}, \mathcal{C}_k)\, \mathrm{d}\mathbf{x}
\end{aligned}
$$

$$
p(\mathcal{C}_k|\mathbf{x})p(\mathbf{x})
$$

Each ***x*** should be assigned the class having the largest posterior probability $p(C_k|\mathbf{x})$

Look Back



x_0
$\widehat{x}$
$p(x, \mathcal{C}_1)$
$p(x, \mathcal{C}_2)$
x
$\mathcal{R}_1$
$\mathcal{R}_2$

Minimum Expected Loss

- In practice, mistakes in predicting different classes may lead to different costs

Example: classify medical images as ‘cancer’ or ‘normal’

Truth \ Decision	cancer	normal
cancer	0	1000
normal	1	0

Minimum Expected Loss

- Define a *cost function*, associate the cost of classifying k to j with L_{kj}

$$\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) \, \mathrm{d}\mathbf{x}.$$

- We want to find the decision regions R_j that minimize the expected loss

Minimum Expected Loss

- Define a *cost function*, associate the cost of classifying k to j with L_{kj}

$$\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) \, \mathrm{d}\mathbf{x}.$$

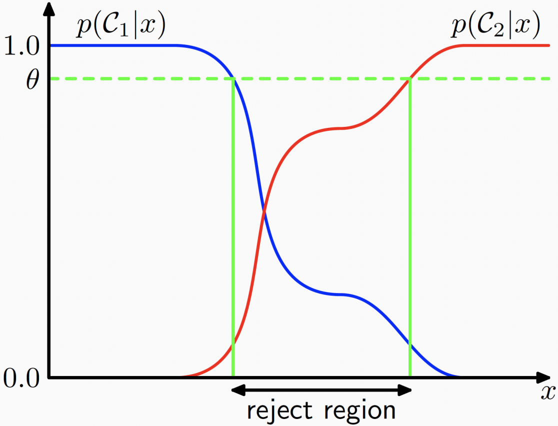
- Rule: Assign each $\boldsymbol{x}$ to the class for which

$$\sum_k L_{kj} p(\mathcal{C}_k | \mathbf{x})$$

is a minimum

Rejct Option

- When the largest posterior probability is still too small



Decision for Continuous Variables

- Inference step
  - Determine $p(\mathbf{x},t)$

- Decision step
  - For any given $\mathbf{x}$, make optimal prediction $y(\mathbf{x})$ for t

Decision for Continuous Variables

- Inference step
  - Determine *p(**x**,t)*

- Decision step
  - For any given ***x***, make optimal prediction *y(**x**)* for t

Loss function

$$\mathbb{E}[L] = \iint L(t, y(\mathbf{x}))p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t$$

The Squared Loss Function

Minimize $\mathbb{E}[L] = \iint \{y(\mathbf{x}) - t\}^2 p(\mathbf{x}, t) \, \mathrm{d}\mathbf{x} \, \mathrm{d}t$

Minimize $\mathbb{E}[L] = \iint \{y(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t$

$$y(\mathbf{x}) = \mathbb{E}[t|\mathbf{x}]$$

What you need to know

- What is the decision? What is the difference between the decision and inference?
- How to find optimal decision regions for classification?
- How to find optimal decisions for continuous variables?

Coding Theory

- Let us start with discrete random variables

Coding Theory

- How to represent the information contained in the random variables?

$$h(\mathbf{x}) \geq 0$$

$$h(\mathbf{x}, \mathbf{y}) = h(\mathbf{x}) + h(\mathbf{y})$$ ***x,y*** are independent

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$$

↓

$$h(\mathbf{x}) = -\log\left(p(\mathbf{x})\right)$$

Entropy

- The average among of information need to transmit

$$H(\mathbf{x}) = -\sum_{\mathbf{x}} p(\mathbf{x}) \log \left(p(\mathbf{x}) \right)$$

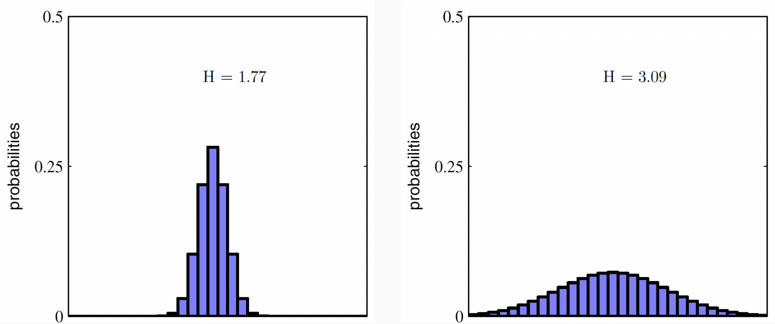
Entropy

x	a	b	c	d	e	f	g	h
$p(x)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{64}$

$$\begin{aligned} \mathrm{H}[x] &= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{16}\log_2\frac{1}{16} - \frac{4}{64}\log_2\frac{1}{64} \\ &= 2 \text{ bits} \end{aligned}$$

Entropy is also the average code length

Entropy Reflects Uncertainty



Maximum Entropy

- Consider a discrete R.V. with M possible status. We want to find the distribution has the the maximum entropy $\mathrm{H}[p] = -\sum_i p(x_i) \ln p(x_i)$.

$$\widetilde{\mathrm{H}} = -\sum_i p(x_i) \ln p(x_i) + \lambda \left(\sum_i p(x_i) - 1 \right)$$

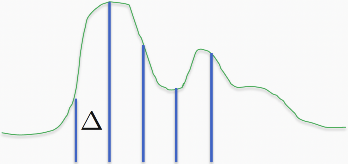
$$p(x_i) = 1/M$$

uniform distribution

- Entropy is naturally defined on discrete random variables.
- But how about continuous variables?

Differential Entropy

- Let us divide x into bins of Δ



Mean-value theorem

$$\int_{i\Delta}^{(i+1)\Delta} p(x)\,\mathrm{d}x = p(x_i)\Delta$$

Entropy on discretized probability

$$\mathrm{H}_\Delta = -\sum_i p(x_i)\Delta \ln\left(p(x_i)\Delta\right) = -\sum_i p(x_i)\Delta \ln p(x_i) - \ln\Delta$$

$$\textstyle\sum_i p(x_i)\Delta = 1$$

Differential Entropy

$$\mathrm{H}_\Delta = -\sum_i p(x_i)\Delta \ln\left(p(x_i)\Delta\right) = -\sum_i p(x_i)\Delta \ln p(x_i) - \ln \Delta$$

Goes to infinity
Throw out it

$$\lim_{\Delta \to 0}\left\{\sum_i p(x_i)\Delta \ln p(x_i)\right\} = \int p(x) \ln p(x)\, \mathrm{d}x$$

$$\mathrm{H}[\mathbf{x}] = -\int p(\mathbf{x}) \ln p(\mathbf{x})\, \mathrm{d}\mathbf{x}$$

Differential Entropy

- The term that is thrown out reflects that to specify a continuous variable very precisely requires many many bits
- Note: differential entropy can be negative!

Differential Entropy

- Given a continuous variable *x* with mean μ and variance σ^2, which distribution has the largest entropy?

$$
\begin{aligned}
\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x &= 1 \\
\int_{-\infty}^{\infty} x p(x)\,\mathrm{d}x &= \mu \\
\int_{-\infty}^{\infty} (x-\mu)^2 p(x)\,\mathrm{d}x &= \sigma^2.
\end{aligned}
$$

Differential Entropy

$$\max \quad -\int_{-\infty}^{\infty} p(x) \ln p(x) \, \mathrm{d}x + \lambda_1 \left(\int_{-\infty}^{\infty} p(x) \, \mathrm{d}x - 1 \right) + \lambda_2 \left(\int_{-\infty}^{\infty} x p(x) \, \mathrm{d}x - \mu \right) + \lambda_3 \left(\int_{-\infty}^{\infty} (x-\mu)^2 p(x) \, \mathrm{d}x - \sigma^2 \right)$$

$$p(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}$$

Gaussian distribution!

Conditional Entropy

- Given ***x***, how much information is left for ***y***

$$\mathrm{H}[\mathbf{y}|\mathbf{x}] = -\iint p(\mathbf{y}, \mathbf{x}) \ln p(\mathbf{y}|\mathbf{x}) \,\mathrm{d}\mathbf{y}\,\mathrm{d}\mathbf{x}$$

$$\mathrm{H}[\mathbf{x}, \mathbf{y}] = \mathrm{H}[\mathbf{y}|\mathbf{x}] + \mathrm{H}[\mathbf{x}]$$

Prove it by yourself

Kullback-Leibler (KL) Divergence

- Also called relative entropy

$$\begin{aligned} \mathrm{KL}(p\|q) &= -\int p(\mathbf{x}) \ln q(\mathbf{x})\, \mathrm{d}\mathbf{x} - \left(-\int p(\mathbf{x}) \ln p(\mathbf{x})\, \mathrm{d}\mathbf{x} \right) \\ &= -\int p(\mathbf{x}) \ln \left\{ \frac{q(\mathbf{x})}{p(\mathbf{x})} \right\} \mathrm{d}\mathbf{x}. \end{aligned}$$

If we use q to transmit information for p, how much extra information do we need

Kullback-Leibler (KL) Divergence

- KL divergence is widely used to measure the difference between two distributions

$$\mathrm{KL}(p||q) \geqslant 0$$

=0 iff p = q

Prove it with convexity
And Jensen's inequality

- However, it is not symmetric!

$$\mathrm{KL}(p||q) \neq \mathrm{KL}(q||p)$$

KL Divergence

- KL divergence plays the key role in approximate inference
- All the deterministic approximate methods aim to minimize the KL divergence between the true and approximate posteriors (or in the reversed direction)
- In general, we have alpha divergence
- We will discuss these in detail later

Mutual Information

How many information do the two random variables share?

$$\mathrm{I}[\mathbf{x}, \mathbf{y}] \equiv \mathrm{KL}(p(\mathbf{x}, \mathbf{y}) \| p(\mathbf{x}) p(\mathbf{y}))$$
$$= -\iint p(\mathbf{x}, \mathbf{y}) \ln \left(\frac{p(\mathbf{x}) p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} \right) \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y}$$

$$\mathrm{I}[\mathbf{x}, \mathbf{y}] = \mathrm{H}[\mathbf{x}] - \mathrm{H}[\mathbf{x}|\mathbf{y}] = \mathrm{H}[\mathbf{y}] - \mathrm{H}[\mathbf{y}|\mathbf{x}]$$

Prove it by yourself

What you need to know

- Definition of entropy
- How is differential entropy is derived
- Entropy is an indicator for uncertainty
- KL divergence and properties (especially asymmetric)
- Mutual information