# CS-567 Machine Learning

Nazar Khan

PUCIT

Lectures 9-13
Nov 10, 12, 17, 19, 24 2015



## Decision Theory

- **Probability Theory**: Mathematical framework for quantifying uncertainty.
- **Decision Theory**: Combines with probability theory to make *optimal decisions* in uncertain scenarios.
- **Inference**: Determining $p(x, t)$ from training data.
- **Decision**: Find a particular $t$.
- $p(x, t)$ is the most complete description of the data.
  - But a decision still needs to be made.
  - This decision is generally very simple after inference.





## Decision Theory
*Example*

- Given X-ray image $\mathbf{x}$, we want to know if the patient has a certain disease or not.
- Let $t = 0$ correspond to the disease class, denoted by $\mathcal{C}_1$.
- Let $t = 1$ correspond to the non-disease class, denoted by $\mathcal{C}_2$.
- Using Bayes' theorem

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

- All quantities can be obtained from $p(\mathbf{x}, t)$ either via marginalization or conditioning.
- **Intuitivley**, to minimise chance of error, assign $\mathbf{x}$ to class with highest posterior.





## Decision Theory

- Any decision rule places inputs $\mathbf{x}$ into *decision regions*.
- If my decision rule places $\mathbf{x}$ in region $\mathcal{R}_1$, I will say that $\mathbf{x}$ belongs to class $\mathcal{C}_1$.
- The probability of $\mathbf{x}$ belonging to class $\mathcal{C}_1$ is $p(\mathbf{x}, \mathcal{C}_1)$. This is the probability of my decision being correct.
- Similarly, the probability of my decision being incorrect is $p(\mathbf{x}, \mathcal{C}_2)$.

## Decision Theory

- When one input $\mathbf{x}$ has been decided upon

$$p(\text{mistake on } \mathbf{x}) = p(\mathbf{x} \text{ placed in region 1 and belongs to class 2} \text{ OR } \mathbf{x} \text{ placed in region 2 and belongs to class 1}) = p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1)$$

- When all inputs have been decided upon

$$p(\text{mistake}) = \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2) d\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1) d\mathbf{x}$$





## Decision Theory

- $p(\text{mistake on } \mathbf{x})$ is minimized when $\mathbf{x}$ is placed in the region $\mathcal{R}_k$ with the highest $p(\mathbf{x}, \mathcal{C}_k)$.
- Overall $p(\text{mistake})$ is minimized when each $\mathbf{x}$ is placed in the region $\mathcal{R}_k$ with the highest $p(\mathbf{x}, \mathcal{C}_k)$.
- Highest $p(\mathbf{x}, \mathcal{C}_k) \implies$ highest $p(\mathcal{C}_k|\mathbf{x})p(\mathbf{x}) \implies$ highest $p(\mathcal{C}_k|\mathbf{x})$.
- For $K$ classes also, $p(\text{mistake})$ is minimised by placing each $\mathbf{x}$ in the region $\mathcal{R}_k$ with highest posterior $p(\mathcal{C}_k|\mathbf{x})$. This is known as the **Bayesian decision rule**.





## Decision Theory
*Loss*

- Suppose we are classifying plant leaves as poisonous or not.
- Are the following mistakes equal?
  - Poisonous leaf classified as non-poisonous.
  - Non-poisonous leaf classified as poisonous.
- We can assign a **loss value** to each mistake.

Classified as

| | poisonous | non-poisonous |
|---|---|---|
| poisonous | 0 | 1000 |
| non-poisonous | 1 | 0 |

- $L_{kj}$ is the loss incurred by classifying a class $k$ item as class $j$.





## Decision Theory
*Loss*

- When mistakes are not equally bad, instead of minimising the **number of mistakes**, it is better to minimize the **expected loss**.

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

- To minimise overall expected loss, place each $\mathbf{x}$ in the region $j$ for which expected loss $\mathbb{E}[L_j]$ is minimum

$$\mathbb{E}[L_j] = \sum_k L_{kj} p(\mathcal{C}_k|\mathbf{x})$$

is minimum.

# Decision Theory

*Reject Option*

- Classification error is high when $p(\mathbf{x}, \mathcal{C}_k)$ (or equivalently $p(\mathcal{C}_k|\mathbf{x})$) is comparable for all $k$.
- Uncertainty because no class is a clear winner.
- **Reject option**: Avoid making a decision on uncertain scenarios.
- Do not make a decision for $\mathbf{x}$ for which largest $p(\mathcal{C}_k|\mathbf{x}) \leq \theta$.
- Loss matrix can include loss of reject option too.

Classified as

| | poisonous | non-poisonous | reject |
|---|---|---|---|
| poisonous | 0 | 1000 | 100 |
| non-poisonous | 1 | 0 | 200 |





# 3 Approaches for Solving Decision Problems

1. **Generative**: Infer posterior $p(\mathcal{C}_k|\mathbf{x})$
   - either by inferring $p(\mathbf{x}|\mathcal{C}_k)$ and $p(\mathbf{x})$ and using Bayes' theorem,
   - or by inferring $p(\mathbf{x}, \mathcal{C}_k)$ and marginalizing.
   - Called generative because $p(\mathbf{x}|\mathcal{C}_k)$ and/or $p(\mathbf{x}, \mathcal{C}_k)$ allow us to generate new $\mathbf{x}$'s.
2. **Discriminative**: Model the posterior $p(\mathcal{C}_k|\mathbf{x})$ directly.
   - If decision depends on posterior, then no need to model the joint distribution.
3. **Discriminant Function**: Just learn a discriminant function that maps $\mathbf{x}$ directly to a class label.
   - $f(\mathbf{x})=0$ for class $\mathcal{C}_1$.
   - $f(\mathbf{x})=1$ for class $\mathcal{C}_2$.
   - No probabilities





# Generative Approach

- For high dimensional $\mathbf{x}$, estimating $p(\mathbf{x}|\mathcal{C}_k)$ requires large training set.
- $p(\mathbf{x})$ allows **outlier detection**. Also called **novelty detection**.
- Estimating $p(\mathcal{C}_k)$ is easy – just use fraction of training data for each class.





# Discriminant Functions

- Directly learn the decision boundaries.
- But now we don't have the posterior probabilities.

## Benefits of knowing the posteriors $p(\mathcal{C}_k|\mathbf{x})$

- If loss matrix changes, decision rule can be trivially revised. Discriminant functions would require retraining.
- Reject option can be used.
- Different models can be combined systematically.





## Combining Models

Let's say we have X-ray images $\mathbf{x}_I$ and blood-tests $\mathbf{x}_B$ and want to classify into disease or not disease.

- **Method 1**: Form $\mathbf{x} = \begin{bmatrix} \mathbf{x}_I \\ \mathbf{x}_B \end{bmatrix}$ and learn classifier for $\mathbf{x}$.
- **Method 2**: Learn $p(\mathcal{C}_k|\mathbf{x}_I)$ and $p(\mathcal{C}_k|\mathbf{x}_B)$.
  - Assuming **conditional independence** $p(\mathbf{x}_I, \mathbf{x}_B|\mathcal{C}_k) = p(\mathbf{x}_I|\mathcal{C}_k)p(\mathbf{x}_B|\mathcal{C}_k)$

$$\begin{aligned} p(\mathcal{C}_k|\mathbf{x}_I, \mathbf{x}_B) &\propto p(\mathbf{x}_I, \mathbf{x}_B|\mathcal{C}_k)p(\mathcal{C}_k) \\ &\propto p(\mathbf{x}_I|\mathcal{C}_k)p(\mathbf{x}_B|\mathcal{C}_k)p(\mathcal{C}_k) \\ &\propto \frac{p(\mathcal{C}_k|\mathbf{x}_I)p(\mathcal{C}_k|\mathbf{x}_B)}{p(\mathcal{C}_k)} \end{aligned}$$

  - Normalise r.h.s using $\sum_k p(\mathcal{C}_k|\mathbf{x}_I, \mathbf{x}_B)$.
  - The conditional independence assumption is also known as the **naive Bayes model**.





## Loss functions for regression

- So far we have used decision theory for classification problems.
- Loss functions can also be defined for regression problems.
- For example, for the polynomial fitting problem a loss function can be described as $L(t, y(\mathbf{x})) = (y(\mathbf{x}) - t)^2$.
- Expected loss can be written as

$$E[L] = \int\int (y(\mathbf{x}) - t)^2 p(\mathbf{x}, t) d\mathbf{x} dt$$

- The minimising polynomial function can be written using calculus of variations as

$$y(\mathbf{x}) = \frac{\int tp(\mathbf{x}, t)dt}{p(\mathbf{x})} = \int tp(t|\mathbf{x})dt = E_t[t|\mathbf{x}]$$

  which is the expected value of $t$ given $\mathbf{x}$. Also called the **regression function**.
- For multivariable outputs $\mathbf{t}$, optimal $y(\mathbf{x}) = E_{\mathbf{t}}[\mathbf{t}|\mathbf{x}]$





## 3 Approaches for Solving Regression Problems

- Similar to the case of classification problems, there are 3 approaches to solve regression problems.
  1. Infer $p(\mathbf{x}, t)$, marginalize to get $p(\mathbf{x})$, normalize to get $p(t|\mathbf{x})$ and use it to compute conditional expectation $E_t[t|\mathbf{x}]$.
  2. Infer $p(t|\mathbf{x})$ directly and use it to compute conditional expectation $E_t[t|\mathbf{x}]$.
  3. Find regression function $y(\mathbf{x})$ directly.
- The relative merits of each approach are similar to those of clasification approaches.

## Information Theory

- Amount of additional information $\propto$ degree of surprise.
- If a highly unlikely event occurs, you gain a lot of new information.
- If an almost certain event occurs, you gain not much new information.
- So information $\propto \frac{1}{\text{probability}}$





## Information Theory

- For unrelated events $x$ and $y$
  - Information from both events should equal information from $x$ plus information from $y$.
  - $p(x, y) = p(x)p(y)$
- From these two relationships, it can be shown that information must be given by the logarithm function.

$$\begin{aligned} h(x, y) &= -\log(p(x, y)) \\ &= -\log(p(x)p(y)) \\ &= -\log(p(x)) - \log(p(y)) \\ h(x) &= -\log(p(x)) \end{aligned}$$

  where $h(x)$ denotes the information given by $x$.
- For base 2 log, units of information $h(x)$ are 'bits'.
- For natural log, units of information $h(x)$ are 'nats' (1 nat$= \ln 2$ bits).





## Information Theory

*Entropy*

- If information given by random variable $x$ is given by a function $h(x) = -\log(p(x))$, then expected information from r.v $x$ is

$$H[x] = E[h(x)] = -\sum \log(p(x))p(x)$$

- Also called the **entropy** of random variable $x$.
- Entropy is just a fancy name for expected information contained in a random variable.





## Information Theory

*Entropy*

- To transmit a r.v $x$ with 8 *equally likely* states, we need 3 bits $(= \log_2 8)$.
- Entropy $H[x] = -\sum \frac{1}{8} \log_2 \frac{1}{8} = 3$ bits.
- For non-uniform probabilities, entropy is reduced.
- **Entropy quantifies order/disorder.**
- Entropy is a lower-bound on the number of bits needed to transmit the state of a random variable.

## Information Theory
*Entropy*

- For a *discrete* r.v $X$ with pdf $p$, entropy is

$$H[p] = -\sum_i p(x_i) \ln p(x_i) \quad (1)$$

- Sharply peaked distribution $\implies$ low entropy.
- Evenly spread distribution $\implies$ high entropy.
- Is the entropy non-negative?
- What is its minimum value?
- When does the minimum value occur?





## Information Theory
*Finding the Maximum Entropy Distribution – Discrete Case*

- How can we find the *discrete* distribution $p(x)$ that maximises the entropy $H[p]$?
- Since $p$ must add up to 1, this a constrained maximisation problem.
- The Lagrangian function is

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

- The maximum is given by the stationary point of $\tilde{H}$.
- Why is it the maximum?





## Information Theory
*Entropy*

- For a *continuous* r.v $X$ with pdf $p$, we define **differential entropy** as

$$H[p] = -\int p(x) \ln p(x) dx$$

- For multivariate $\mathbf{x}$

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





## Information Theory
*Finding the Maximum Entropy Distribution – Discrete Case*

- How can we find the *continuous* distribution $p(x)$ that maximises the entropy $H[p]$?
- The maximum entropy discrete distribution was the **uniform** distribution.
- The maximum differential entropy continuous distribution is the **Gaussian** distribution (Excercise 1.34 in Bishop's book).

## Information Theory
*Entropy*

- Differential entropy of the Gaussian is

$$H[x] = \frac{1}{2}\{1 + \ln(2\pi\sigma^2)\}$$

- Proportional to $\sigma^2$. Entropy increases as more values become probable.
- Can also be negative (for $\sigma^2 < \frac{1}{2\pi e}$).





## Information Theory
*Conditional Entropy*

- Let $p(\mathbf{x}, \mathbf{y})$ be a joint distribution.
- Given $\mathbf{x}$, additional information needed to specify $\mathbf{y}$ is the conditional information $-\ln(p(\mathbf{y}|\mathbf{x}))$.
- So expected conditional information is

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

- Also called the **conditional entropy** of $\mathbf{y}$ given $\mathbf{x}$.
- Satisfies $H[\mathbf{x}, \mathbf{y}] = H[\mathbf{y}|\mathbf{x}] + H[\mathbf{x}]$. Information needed to specify $\mathbf{x}$ and $\mathbf{y}$ equals information for $\mathbf{x}$ alone plus *additional* information needed to specify $\mathbf{y}$ given $\mathbf{x}$.





## Information Theory
*Relative entropy*

- Let r.v. $\mathbf{x}$ have a true distribution $p(\mathbf{x})$ and let our estimate of this distribution be $q(\mathbf{x})$.
- Average information required to specify $x$ when its information content is determined using $p(\mathbf{x})$ is given by the entropy

$$H[p] = -\int p(\mathbf{x}) \ln p(\mathbf{x}) \tag{2}$$

- Average information required to specify $x$ when its information content is determined using $q(\mathbf{x})$ is given by

$$\tilde{H}[q] = -\int p(\mathbf{x}) \ln q(\mathbf{x}) \tag{3}$$





## Information Theory
*Relative entropy*

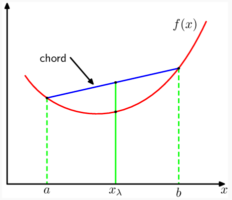
- Average *additional* information required to specify $x$ when $q(\mathbf{x})$ is used instead of $p(\mathbf{x})$ is given by $\tilde{H}[q] - H[p] = (-\int p(\mathbf{x}) \ln q(\mathbf{x})) - (-\int p(\mathbf{x}) \ln p(\mathbf{x}))$.
- This is known as the **relative entropy**, or **Kullback-Leibler (KL) divergence**.

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

- $KL(p||q) \neq KL(q||p)$.
- $KL(p||q) \geq 0$ with equality for $p = q$.

## Convex Functions

- A function $f(x)$ is **convex** if every chord lies on or above the function.
- Any value of $x$ in the interval $a$ to $b$ can be parameterised as $\lambda a + (1-\lambda)b$ where $0 \le \lambda \le 1$.
- The corresponding point on the chord can be parameterised as $\lambda f(a) + (1-\lambda) f(b)$.
- The corresponding point on the function can be parameterised as $f(\lambda a + (1-\lambda)b)$.








## Convex Functions

- Convexity implies points on chord lie on or above points on function. That is

$$f(\lambda a + (1-\lambda)b) \le \lambda f(a) + (1-\lambda) f(b)$$

- Convexity is equivalent to positive second derivative everywhere.
- If function and chord are equal only for $\lambda = 0$ and $\lambda = 1$, then the function is called **strictly convex**.
- The inverse property (every chord lies on or below the function) is called **concavity**.
- If $f(x)$ is convex, then $-f(x)$ will be concave.





## Jensen's Inequality

- Every convex function $f(x)$ satisfies the so-called **Jensen's inequality**

$$f\left(\sum_{i=1}^{M} \lambda_i x_i\right) \le \sum_{i=1}^{M} \lambda_i f(x_i)$$

  where $\lambda_i \ge 0$ and $\sum_{i=1}^{M} \lambda_i = 1$ for any set of points $(x_1, \ldots, x_M)$.
- Interpreting the $\lambda_i$ as probabilities $p(x_i)$, Jensen's inequality can be formulated for *discrete random variables* as

$$f(\mathbb{E}[x]) \le \mathbb{E}[f(x)]$$

- For *continuous random variables*, Jensen's inequality becomes

$$f\left(\int \mathbf{x} p(\mathbf{x} d\mathbf{x}\right) \le \int f(\mathbf{x})\, p(\mathbf{x} d\mathbf{x}$$





## KL-divergence

- Using Jensen's inequality

$$KL(p||q) = -\int p(\mathbf{x}) \underbrace{\ln\left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\}}_{\text{concave}} d\mathbf{x} \ge \underbrace{-\ln \underbrace{\int q(\mathbf{x}) d\mathbf{x}}_{=1}}_{=0}$$

  where the equality holds only when $p(\mathbf{x}) = q(\mathbf{x})\ \forall \mathbf{x}$ (because $-\ln x$ is strictly convex).
- Since $KL(p||q) \ge 0$ and $KL(p||p) = 0$, KL-divergence can be interpreted as a **measure of dissimilarity** between distributions $p(\mathbf{x})$ and $q(\mathbf{x})$.

## Relation between data compression and density estimation

- Optimal compression requires the true density.
- For estimated density, KL-divergence gives **average, additional information** required by **transmitting via estimated density** instead of true density.





## Density Estimation via KL-divergence

- Suppose we have finite data points $\mathbf{x}_1, \dots, \mathbf{x}_N$ drawn from an *unknown* distribution $p(\mathbf{x})$.
- We want to approximate $p(\mathbf{x})$ by some parametric distribution $q(\mathbf{x}|\boldsymbol{\theta})$.
- We can do this by finding $\boldsymbol{\theta}$ that minimizes $KL(p||q)$. **But** $p$ **is unknown.**
- However, $KL(p||q)$ is an *expectation w.r.t* $p(\mathbf{x})$ and can be approximated by the ordinary average for large $N$ (law of large numbers). So

$$\begin{aligned} KL(p||q) &= -\int p(\mathbf{x}) \ln \left\{ \frac{q(\mathbf{x}|\boldsymbol{\theta})}{p(\mathbf{x})} \right\} d\mathbf{x} \qquad (4) \\ &\approx \frac{1}{N} \sum_{n=1}^{N} \{-\ln q(\mathbf{x}_n|\boldsymbol{\theta}) + \ln p(\mathbf{x})\} \end{aligned}$$





## Density Estimation via KL-divergence

- Minimizing w.r.t $\boldsymbol{\theta}$ is equivalent to minimizing $\sum_{n=1}^{N} -\ln q(\mathbf{x}_n|\boldsymbol{\theta})$ which is the **negative log-likelihood of data under** $q(\mathbf{x}|\boldsymbol{\theta})$.
- So *minimizing KL-divergence is equivalent to maximising likelihood (ML estimation)*.





## Mutual Information

- Given 2 random variables $\mathbf{x}$ and $\mathbf{y}$, can we find *how independent* they are?
- If they are independent then $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$. So $KL(p(\mathbf{x}, \mathbf{y})||p(\mathbf{x})p(\mathbf{y})) = 0$.
- Therefore, $KL(p(\mathbf{x}, \mathbf{y})||p(\mathbf{x})p(\mathbf{y}))$ is a measure of *how independent* $\mathbf{x}$ and $\mathbf{y}$ are.
- Also called the **mutual information** $I[\mathbf{x}, \mathbf{y}]$ between variables $\mathbf{x}$ and $\mathbf{y}$.

$$\begin{aligned} I[\mathbf{x}, \mathbf{y}] &= KL(p(\mathbf{x}, \mathbf{y})||p(\mathbf{x})p(\mathbf{y})) \qquad (5) \\ &= -\int\int p(\mathbf{x}, \mathbf{y}) \ln \left( \frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} \right) d\mathbf{x} d\mathbf{y} \end{aligned}$$

- $I[\mathbf{x}, \mathbf{y}] \geq 0$ with equality iff $\mathbf{x}$ and $\mathbf{y}$ are independent.

## Mutual Information

- Using the sum and product rules

$$I[\mathrm{x},\mathrm{y}] = \underbrace{H[\mathrm{x}]}_{\substack{\text{avg. info. needed} \\ \text{to transmit } \mathbf{x}}} - \underbrace{H[\mathrm{x}|\mathrm{y}]}_{\substack{\text{avg. info. needed} \\ \text{to transmit } \mathbf{x} \\ \text{knowing state of } \mathbf{y}}}$$

$$= \underbrace{H[\mathrm{y}]}_{\substack{\text{avg. info. needed} \\ \text{to transmit } \mathbf{y}}} - \underbrace{H[\mathrm{y}|\mathrm{x}]}_{\substack{\text{avg. info. needed} \\ \text{to transmit } \mathbf{y} \\ \text{knowing state of } \mathbf{x}}}$$

- Mutual information captures
  - Information about **x** that is contained in **y**.
  - Information about **y** that is contained in **x**.
  - Reduction in uncertainty of one variable when the other is known.