

CS-567 Machine Learning

Nazar Khan

PUCIT

Lectures 9-13

Nov 24, 26, December 1, 2, 3 2014

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.
- ▶ **Intuitively**, 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*.
- ▶ Assign all $\mathbf{x}$ falling in decision region $\mathcal{R}_k$ to class $\mathcal{C}_k$.
- ▶ Mistake occurs when $\mathbf{x}$ lies in region k but is assigned to class $j \neq k$.

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

Decision Theory

- ▶ $p(\text{mistake})$ is minimized by assigning each $\mathbf{x}$ to the class $\mathcal{C}_k$ with 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, error is minimised by assigning each $\mathbf{x}$ to the class $\mathcal{C}_k$ with highest posterior $p(\mathcal{C}_k|\mathbf{x})$.

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

- ▶ Instead of minimising the mistakes, it is better to minimize the **expected loss**.

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

- ▶ To minimise expected loss, assign each $\mathbf{x}$ to the class j for which

$$\sum_k L_{kj} p(C_k | \mathbf{x})$$

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.

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 $p(C_k|\mathbf{x})$

- ▶ 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 $p(C_k|\mathbf{x}_I)$ and $p(C_k|\mathbf{x}_B)$.
 - ▶ Assuming **conditional independence**
 $p(\mathbf{x}_I, \mathbf{x}_B|C_k) = p(\mathbf{x}_I|C_k)p(\mathbf{x}_B|C_k)$

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

- ▶ Normalise r.h.s using $\sum_k p(C_k|\mathbf{x}_I, \mathbf{x}_B)$.
- ▶ 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

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

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 classification 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 r.v 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 r.v x .

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 (Exercise 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 σ^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. x have a true distribution $p(x)$ and let our estimate of this distribution be $q(x)$.
- ▶ Average information required to specify x when its information content is determined using $p(x)$ is given by the entropy

$$H[p] = - \int p(x) \ln p(x) \quad (2)$$

- ▶ Average information required to specify x when its information content is determined using $q(x)$ is given by

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

Information Theory

Relative entropy

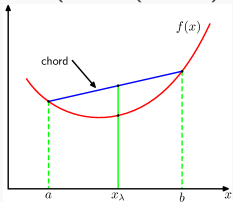
- ▶ Average *additional* information required to specify x when $q(x)$ is used instead of $p(x)$ is given by
$$\tilde{H}[q] - H[p] = \left(-\int p(x) \ln q(x)\right) - \left(-\int p(x) \ln p(x)\right).$$
- ▶ This is known as the **relative entropy**, or **Kullback-Leibler (KL) divergence**.

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

- ▶ $KL(p||q) \neq KL(q||p)$.
- ▶ $KL(p||q) \geq 0$ 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 $0 \leq \lambda \leq 1$.
- ▶ 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

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

- ▶ 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**

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

where $\lambda_i \geq 0$ and $\sum_{i=1}^M \lambda_i = 1$ for any set of points $(x_1, \dots, x_M)$.

- ▶ Interpreting the λ_i as probabilities $p(x_i)$, Jensen's inequality can be formulated for *discrete random variables* as

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

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

$$f\left(\int x p(x) dx\right) \leq \int f(x) p(x) dx$$

KL-divergence

- ▶ Using Jensen's inequality

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

where the equality holds only when $p(x) = q(x) \forall x$ (because $-\ln x$ is strictly convex).

- ▶ Since $KL(p||q) \geq 0$ and $KL(p||p) = 0$, KL-divergence can be interpreted as a **measure of dissimilarity** between distributions $p(x)$ and $q(x)$.

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

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

Density Estimation via KL-divergence

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

Mutual Information

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

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

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

Mutual Information

- ▶ Using the sum and product rules

$$\begin{aligned} I[x, y] &= \underbrace{H[x]}_{\text{avg. info. needed to transmit } x} - \underbrace{H[x|y]}_{\text{avg. info. needed to transmit } x \text{ knowing state of } y} \\ &= \underbrace{H[y]}_{\text{avg. info. needed to transmit } y} - \underbrace{H[y|x]}_{\text{avg. info. needed to transmit } y \text{ knowing state of } x} \end{aligned}$$

- ▶ 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.