CS-567 Machine Learning

Nazar Khan

PUCIT

Lecture 8
Decision theory

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.

In this lecture . . .

- Decisions to ensure minimum misclassifications.
- Decisions to ensure minimum loss.
- Decisions with multiple models.
- Generative vs. discriminative vs. discriminant function approaches.
- Decision theory for regression.

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.

Minimizing Misclassifications

- 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 $p(\mathbf{x}, \mathcal{C}_2)$.

Minimizing Misclassifications

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

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

- When all inputs have been decided upon

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

Minimizing Misclassifications

- Individual $p(\text{mistake on } \mathbf{x})$ 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**.

Minimizing 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.

$$
\begin{array}{cc}
 & \text{Classified as} \\
\begin{array}{l} \\ \text{poisonous} \\ \text{non-poisonous} \end{array} &
\begin{pmatrix} \text{poisonous} & \text{non-poisonous} \\ 0 & 1000 \\ 1 & 0 \end{pmatrix}
\end{array}
$$

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

Minimizing Loss
Finding the optimal decision rule

- Let $\mathcal{R}_j$ consist of all points assigned to class $\mathcal{C}_j$.
- The loss of assigning a point belonging to class $\mathcal{C}_k$ to the class $\mathcal{C}_j$ is denoted by L_{kj}.
- Probability of points *assigned* to class $\mathcal{C}_j$ *belonging* to class $\mathcal{C}_k$ can be written as

$$\int_{\mathcal{R}_j} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x} \text{ Why?}$$

- Note that we do not know which class any $\mathbf{x}$ belongs to. So we are using the probabilities of belonging to each class.

Minimizing Loss

Finding the optimal decision rule

- Expected loss $\mathbb{E}[L_{kj}]$ of assigning points belonging to $\mathcal{C}_k$ to class $\mathcal{C}_j$ can be written as

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

- Overall expected loss due to misclassifications can be written as

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

- Expected loss of assigning a new point $\mathbf{x}$ to class $\mathcal{C}_j$ can be written as

$$\sum_k L_{kj} p(\mathbf{x}, \mathcal{C}_k) \text{ Why?}$$

Minimizing Loss

Finding the optimal decision rule

- So to minimise the overall expected loss, *assign each $\mathbf{x}$ to the class $\mathcal{C}_j$ for which expected loss $\sum_k L_{kj} p(\mathbf{x}, \mathcal{C}_k)$ is minimum*.
- Since $p(\mathbf{x}, C_k) \propto p(C_k|\mathbf{x})$, this rule is the same as assigning each $\mathbf{x}$ to the class $\mathcal{C}_j$ for which expected loss $\sum_k L_{kj} p(\mathcal{C}_k|\mathbf{x})$ is minimum

Minimizing Loss
Summary

- 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.

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 for 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.

$$
\begin{array}{l c}
 & \text{Classified as} \\
 & \begin{array}{ccc} \text{poisonous} & \text{non-poisonous} & \text{reject} \end{array} \\
\begin{array}{l} \text{poisonous} \\ \text{non-poisonous} \end{array} & \begin{pmatrix} 0 & 1000 & 100 \\ 1 & 0 & 200 \end{pmatrix}
\end{array}
$$

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 classification approaches.

Summary

- Decision rule to ensure minimum misclassifications is to assign to class with highest posterior $p(C_k|\mathbf{x})$.
- Decision rule to ensure minimum loss is to assign to class with lowest expected loss.
- Though they are derived mathematically, both are common sense rules.
- Reject option can be used for highly uncertain scenarios.
- Multiple models can be combined via Naive Bayes assumption.
- Generative vs. discriminative vs. discriminant function approaches.
- Decision theory for regression problems leads to similar conclusions as classification.