Decision Theory

Sargur Srihari srihari@cedar.buffalo.edu

Decision Theory

- Using probability theory to make optimal decisions
- Input vector x, target vector t
 - Regression: t is continuous
 - Classification: t will consist of class labels
- Summary of uncertainty associated is given by p(x,t)
- Inference problem is to obtain p(x,t) from data
- Decision: make specific prediction for value of t and take specific actions based on t

Medical Diagnosis Problem

- X-ray image of patient
- Whether patient has cancer or not
- Input vector x is set of pixel intensities
- Output variable t represents whether cancer or not C₁ is cancer and C₂ is absence of cancer
- General inference problem is to determine $p(x,C_k)$ which gives most complete description of situation
- In the end we need to decide whether to give treatment or not. Decision theory helps do this

Bayes Decision

- How do probabilities play a role in making a decision?
- Given input x and classes C_k using Bayes theorem

$$p(C_k \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C_k)p(C_k)}{p(\mathbf{x})}$$

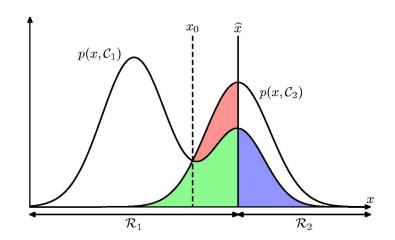
 Quantities in Bayes theorem can be obtained from p(x,C_k) either by marginalizing or conditioning wrt appropriate variable

Minimizing Expected Error

Probability of mistake (2-class)

$$P(error) = p(x \varepsilon R_1, C_2) + p(x \varepsilon R_2, C_1)$$
$$= \int_{R_1} p(x, C_2) dx + \int_{R_2} p(x, C_1) dx$$

- Minimum error decision rule
 - For a given x choose class for which integrand is smaller
 - Since $p(x,C_k)=p(C_k|x)p(x)$, choose class for which a posteriori probability is highest
 - Called Bayes Classifier



Single input variable *x*

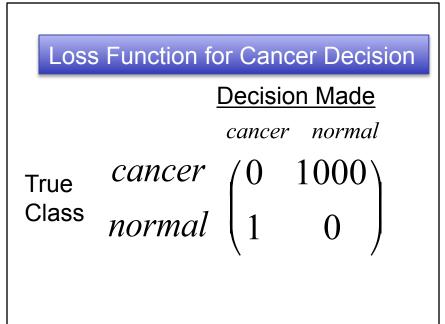
If priors are equal, decision is based on class-conditional densities $p(x|C_k)$

Minimizing Expected Loss

- Unequal importance of mistakes
- Medical Diagnosis
- Loss or Cost Function given by Loss Matrix
- Utility is negative of Loss
- Minimize Average Loss

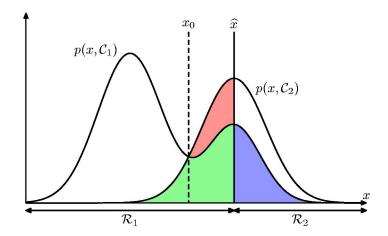
$$E[L] = \sum_{k} \sum_{j} \int_{R_{j}} L_{kj} p(\mathbf{x}, C_{k}) d\mathbf{x}$$

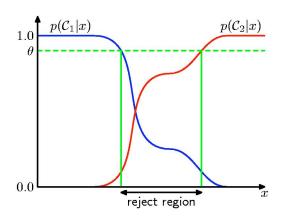
- Minimum Loss Decision Rule
 - Choose class for which $\sum_{k} L_{kj} p(C_k \mid \mathbf{x})$ is minimum
 - Trivial once we know a posteriori probabilities



Reject Option

- Decisions can be made when a posteriori probabilities are significantly less than unity or joint probabilities have comparable values
- Avoid making decisions on difficult cases





Inference and Decision

- Classification problem broken into two separate stages
 - Inference, where training data is used to learn a model for $p(C_k,x)$
 - Decision, use posterior probabilities to make optimal class assignments
- Alternatively can learn a function that maps inputs directly into labels
- Three distinct approaches to Decision Problems
 - 1. Generative
 - 2. Discriminative
 - 3. Discriminant Function

1. Generative Models

- First solve inference problem of determining class-conditional densities $p(\mathbf{x}|C_k)$ for each class separately
- Then use Bayes theorem to determine posterior probabilities

$$p(C_k \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C_k)p(C_k)}{p(\mathbf{x})}$$

 Then use decision theory to determine class membership

2. Discriminative Models

• First solve inference problem to determine posterior class probabilities $p(C_k|x)$

Use decision theory to determine class membership

3. Discriminant Functions

- Find a function f (x) that maps each input x directly to class label
 - In two-class problem, f (.) is binary valued
 - f=0 represents class C_1 and f=1 represents class C_2
- Probabilities play no role
 - No access to posterior probabilities $p(C_k|\mathbf{x})$

Need for Posterior Probabilities

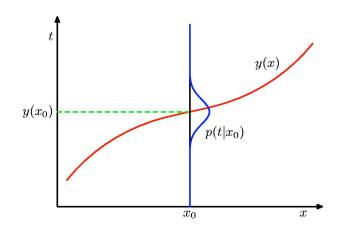
- Minimizing risk
 - Loss matrix may be revised periodically as in a financial application
- Reject option
 - Minimize misclassification rate, or expected loss for a given fraction of rejected points
- Compensating for class priors
 - When far more samples from one class compared to another, we use a balanced data set (otherwise we may have 99.9% accuracy always classifying into one class)
 - Take posterior probabilities from balanced data set, divide by class fractions in the data set and multiply by class fractions in population to which the model is applied
 - Cannot be done if posterior probabilities are unavailable
- Combining models
 - X-ray images (x_I) and Blood tests (x_R)
 - When posterior probabilities are available they can be combined using rules of probability
 - Assume feature independence $p(x_l, x_B|C_k) = p(x_l, C_k) p(x_B, C_k)$ [Naïve Bayes Assumption]
 - Then $p(C_k|\mathbf{x}_I,\mathbf{x}_B) \ \alpha \quad p(\mathbf{x}_I,\mathbf{x}_B|C_k)p(C_k)$ $\alpha \quad p(\mathbf{x}_I,|C_k) \ p(\mathbf{x}_B,|C_k) \ p(C_k)$ $\alpha \quad p(C_k|\mathbf{x}_I) \ p(C_k|\mathbf{x}_B)/p(C_k)$
 - Need $p(C_k)$ which can be determined from fraction of data points in each class. Then need to normalize resulting probabilities to sum to one

Loss Functions for Regression

- Curve fitting can also use a loss function
- Regression decision is to choose a specific estimate y(x) of t for a given x
- Incur loss L(t,y(x))

 $y(x) = E_t[t|x]$

- Squared loss function $L(t,y(x)) = \{y(x)-t\}^2$
- Minimize expected loss
 E[L] = ∫∫ L(t,y(x))p(x,t)dxdt
 Taking derivative and setting equal to zero yields a solution



Regression function y(x), which minimizes the expected squared loss, is given by the mean of the conditional distribution p(t|x)

Inference and Decision for Regression

- Three distinct approaches (decreasing complexity)
- Analogous to those for classifiction
 - 1. Determine joint density $p(\mathbf{x},t)$ Then normalize to find conditional density $p(t|\mathbf{x})$ Finally marginalize to find conditional mean $E_t[t|\mathbf{x}]$
 - Solve inference problem of determining conditional density p(t|x)
 Marginalize to find conditional mean
 - 3. Find regression function y(x) directly from training data

Minkowski Loss Function

- Squared Loss is not only possible choice for regression
- Important example concerns multimodal $p(t|\mathbf{x})$
- Minkowski Loss $L_q = |y-t|^q$
- Minimum of E[t|x] is given by
 - conditional mean for q=2,
 - conditional median for q=1 and
 - conditional mode for $q \rightarrow 0$

