CSE 4621
Machine Learning

Lecture 11

Md. Hasanul Kabir, PhD.

Professor, CSE Department

Islamic University of Technology (IUT)

Probabilistic Classification

- Probabilistic classification means that the model used for classification is a probabilistic model.
- Probabilistic model can give probability of an instance belonging to positive or negative class. Then it is up to us to decide whether the instance is positive or negative based on the probabilities given by the model.

Probabilistic Classification

Input: $S_{\text{train}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ training examples

$$y_i \in \{c_1, c_2, \ldots, c_J\}$$

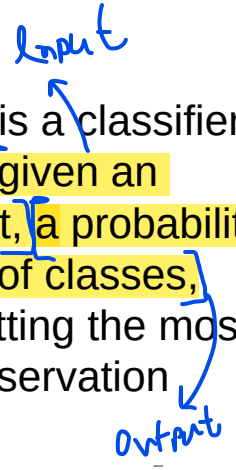
Goal: $h : X \to Y$
For each class c_j, estimate

$$P(y = c_j \mid \boldsymbol{x}, S_{\text{train}})$$

Assign to $\boldsymbol{x}$ the class with the highest probability

$$\hat{y} = h(\boldsymbol{x}) = \arg\max_{c} P(y = c \mid \boldsymbol{x}, S_{\text{train}})$$

- Probabilistic classifier is a classifier that is able to predict, given an observation of an input, a probability distribution over a set of classes, rather than only outputting the most likely class that the observation should belong to.



Decision Theory

- Probability theory provides us with a consistent mathematical framework for quantifying and manipulating uncertainty.
- **Decision theory**: When combined with probability theory, that allows us to make optimal decisions in situations involving uncertainty.
- Suppose we have an input vector **x** together with a corresponding vector **t** of target variables, and our goal is to predict **t** given a new value for **x**.
  - For regression problems, **t** will comprise continuous variables
  - For classification problems **t or C_k** will represent class labels.
- The joint probability distribution $p(\mathbf{x}, \mathbf{t})$ or $p(x, C_k)$, provides a complete summary of the uncertainty associated with these variables.
- Determination of $p(\mathbf{x}, \mathbf{t})$ from a set of training data is an example of *inference*
  - is typically a very difficult problem
- Finally, the *decision* step, it is the subject of decision theory to tell us how to make optimal decisions given the appropriate probabilities.

Example Scenario

- When we obtain the X-ray image **x** for a new patient, our goal is to decide which of the two classes to assign to the image. We are interested in the probabilities of the two classes given the image, which are given by $p(C_k/\mathbf{x})$.
- Using Bayes' theorem, these probabilities can be expressed in the form

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

- Interpret $p(C_k)$ as the prior probability for the class C_k, and $p(C_k/\mathbf{x})$ as the corresponding posterior probability. Thus $p(C_1)$ represents the probability that a person has cancer, before we take the X-ray measurement.
- If our aim is to minimize the chance of assigning **x** to the wrong class, then intuitively we would choose the class having the higher posterior probability.

Approaches to Solving Decision Problems

- Two separate stages in Decision Theory,
  - the *inference stage* in which we use training data to learn a model for $p(C_k|\mathbf{x})$, and
  - the subsequent *decision* stage in which we use these posterior probabilities to make optimal class assignments.
    - Generative Model
    - Discriminative Model
- An alternative possibility would be to solve both problems together and simply learn a function that maps inputs **x** directly into decisions. Such a function is called a *discriminant function*.
  - Discriminant Function

(Compare between them)

Generative Model

- First solve the inference problem of determining the class-conditional densities $p(\mathbf{x}|C_k)$ for each class C_k individually. Also separately infer the prior class probabilities $p(C_k)$. Then use Bayes' theorem in the form

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

- Equivalently, we can model the joint distribution $p(\mathbf{x}, C_k)$ directly and then normalize to obtain the posterior probabilities.
- Approaches that explicitly or implicitly model the distribution of inputs as well as outputs are known as *generative models*,
  - by sampling from them it is possible to generate synthetic data points in the input space.

Examples of Generative Models

- Gaussian mixture model (and other types of mixture model)
- Hidden Markov model
- Probabilistic context-free grammar
- Bayesian network (e.g. **Naive Bayes**, Autoregressive model)
- Averaged one-dependence estimators
- Latent Dirichlet allocation
- Boltzmann machine (e.g. Restricted Boltzmann machine, Deep belief network)
- Variational autoencoder
- Generative adversarial network
- Flow-based generative model
- Energy based model

Discriminative Model

- First solve the inference problem of determining the posterior class probabilities $p(C_k|\mathbf{x})$, and then subsequently use decision theory to assign each new **x** to one of the classes.
- Approaches that model the posterior probabilities directly are called *discriminative models*.

- Examples:

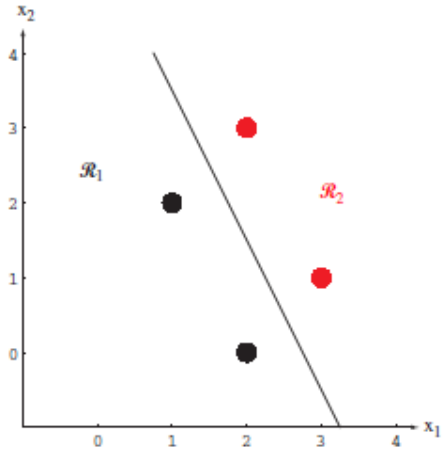
  k-nearest neighbors algorithm, Logistic regression, Support Vector Machines, Decision Trees, Random Forest, Maximum-entropy Markov models, Conditional random fields, **Neural networks**

Discriminant Function

- Find a function $f(\mathbf{x})$, called a discriminant function, which maps each input $\mathbf{x}$ directly onto a class label.
- For instance, in the case of two-class problems, $f(\cdot)$ might be binary valued and such that $f = 0$ represents class $C1$ and $f = 1$ represents class $C2$. In this case, probabilities play no role.

The linear discriminant function $g(\mathbf{x})$ can be written as

$$g(\mathbf{x}) = w_0 + \sum_{i=1}^{d} w_i x_i,$$



Graphical Model

- *Graphical models*, also called *Bayesian networks*, *belief networks*, or *probabilistic networks*, are composed of nodes and arcs between the nodes.
- Each node corresponds to a random variable, *X*, and has a value corresponding to the probability of the random variable, *P(X)*. If there is a directed arc from node *X* to node *Y*, this indicates that *X* has a *direct influence* on *Y*. This influence is specified by the conditional probability *P(Y|X)*. The network is a *directed acyclic graph* (DAG)
- The nodes and the arcs between the nodes define the *structure* of the network, and the conditional probabilities are the *parameters* given the structure.
- Graphical models are frequently used to visualize *generative models* for representing the process that we believe has created the data.

Example

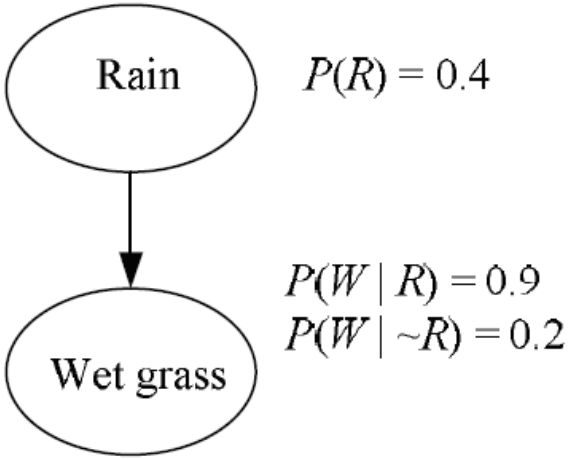
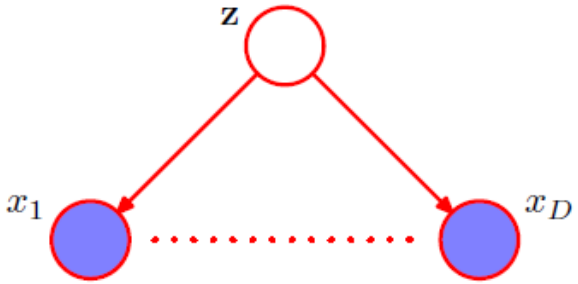


Figure 14.1 Bayesian network modeling that rain is the cause of wet grass.

Naïve Bayes Classifier

- **Naive Bayes classifier is a** probabilistic classifier based on applying Bayes' theorem with strong (naïve) independence assumptions between the features.
- In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.
- Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to sometimes outperform even highly sophisticated classification methods.



A graphical representation of the 'naive Bayes' model for classification. Conditioned on the class label $\mathbf{z}$, the components of the observed vector $\mathbf{x} = (x_1, \ldots, x_D)^{\mathrm{T}}$ are assumed to be independent.

Applications of Naïve Bayes Classifier

- **Disease Prediction:** Naive Bayes is an eager learning classifier and it is sure fast. Thus, it could be used for making predictions in real time.
- **Multi class Prediction:** This algorithm is also well known for multi class prediction feature. Here we can predict the probability of multiple classes of target variable.
- **Text classification/ Spam Filtering/ Sentiment Analysis:** Naive Bayes classifiers mostly used in text classification (due to better result in multi class problems and independence rule) have higher success rate as compared to other algorithms. As a result, it is widely used in Spam filtering (identify spam e-mail) and Sentiment Analysis (in social media analysis, to identify positive and negative customer sentiments)
- **Recommendation System:** Naive Bayes Classifier and Collaborative Filtering together builds a Recommendation System that uses machine learning and data mining techniques to filter unseen information and predict whether a user would like a given resource or not.

How it works

- Naïve Bayes is a conditional probability model: given a unknown sample to be classified, represented by a vector x=(x_1,x_2,…,x*n*) representing some *n* features (independent variables), it assigns to this posterior probabilities

$$P(C_K \mid x_1, x_2, .., x_n) = P(C_K \mid x)$$

for each of K possible outcomes or classes C_K.

- Bayes theorem provides a way of calculating posterior probability $P(C_k|x)$ from $P(C_K)$, P(x) and $P(x|C_k)$

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

- In practice, there is interest only in the numerator of that fraction. The numerator is equivalent to the joint probability model

$$P(C_K, x_1, x_2, .., x_n)$$

How it works

$$p(C_k, x_1, \dots, x_n)$$

which can be rewritten as follows, using the chain rule for repeated applications of the definition of conditional probability:

$$\begin{aligned} p(C_k, x_1, \dots, x_n) &= p(x_1, \dots, x_n, C_k) \\ &= p(x_1 \mid x_2, \dots, x_n, C_k)\, p(x_2, \dots, x_n, C_k) \\ &= p(x_1 \mid x_2, \dots, x_n, C_k)\, p(x_2 \mid x_3, \dots, x_n, C_k)\, p(x_3, \dots, x_n, C_k) \\ &= \cdots \\ &= p(x_1 \mid x_2, \dots, x_n, C_k)\, p(x_2 \mid x_3, \dots, x_n, C_k) \cdots p(x_{n-1} \mid x_n, C_k)\, p(x_n \mid C_k)\, p(C_k) \end{aligned}$$

Now the "naïve" conditional independence assumptions come into play: assume that all features in $\mathbf{x}$ are mutually independent, conditional on the category C

Under this assumption,

$$p(x_i \mid x_{i+1}, \dots, x_n, C_k) = p(x_i \mid C_k)$$

Thus, the joint model can be expressed as

$$\begin{aligned} p(C_k \mid x_1, \dots, x_n) &\propto p(C_k, x_1, \dots, x_n) \\ &\propto p(C_k)\, p(x_1 \mid C_k)\, p(x_2 \mid C_k)\, p(x_3 \mid C_k) \cdots \\ &\propto p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k), \end{aligned}$$

where $\propto$ denotes proportionality.

The naïve Bayes classifier combines this model with a decision rule. One common rule is to pick the hypothesis that is most probable; this is known as the maximum a posteriori or MAP decision rule.

$$\hat{y} = \underset{k \in \{1,2,..,K\}}{\arg\max}\; p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)$$

Example

Outlook

	Yes	No	P(yes)	P(no)
Sunny	2	3	2/9	3/5
Overcast	4	0	4/9	0/5
Rainy	3	2	3/9	2/5
Total	9	5	100%	100%

Temperature

	Yes	No	P(yes)	P(no)
Hot	2	2	2/9	2/5
Mild	4	2	4/9	2/5
Cool	3	1	3/9	1/5
Total	9	5	100%	100%

Humidity

	Yes	No	P(yes)	P(no)
High	3	4	3/9	4/5
Normal	6	1	6/9	1/5
Total	9	5	100%	100%

Wind

	Yes	No	P(yes)	P(no)
False	6	2	6/9	2/5
True	3	3	3/9	3/5
Total	9	5	100%	100%

Play		P(Yes)/P(No)
Yes	9	9/14
No	5	5/14
Total	14	100%

Example

- Today = (Outlook = **Sunny**, Temperature = **Hot**, Humidity = **Normal**, Wind = **False**)
- So, probability of playing golf (YES) is given by:

- probability to not play golf (NO) is given by:

Pros:

- It is easy and fast to predict class of test data set. It also perform well in multi class prediction
- When assumption of independence holds, a Naive Bayes classifier performs better compare to other models like logistic regression and you need less training data.
- It perform well in case of categorical input variables compared to numerical variable(s). For numerical variable, normal distribution is assumed (bell curve, which is a strong assumption).

Cons:

- *If categorical variable has a category (in test data set), which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as "Zero Frequency".*
- *On the other side naive Bayes is also known as a bad estimator, so the probability outputs are not to be taken too seriously.*
- *Assumption of independent predictors. In real life, it is almost impossible that we get a set of predictors which are completely independent.*