Basic Concepts in Machine Learning

Pengyu Hong

Applications

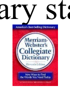
- Vision (e.g., Face/Gesture recognition, …)
- Speech (e.g., Phoneme/Word/Speaker identification, …)
- Data mining (e.g., Association rules, …)
- Time-series prediction (e.g., Traffic/Asset prediction, …)
- Natural language processing (e.g., Parsing/Translation/POS, …)
- Bioinformatics and computational biology (e.g., gene detection/prediction, disease diagnosis, biological network modeling, …)
- Medical Informatics (evidence-based, data-intensive healthcare, …)
- … …

- Probability – calculus of uncertainty for making assessment and prediction.

- Statistics – rigorously study and quantify the characteristics of a sub-population and draw a conclusion appropriate for the entire population.

Random Variables

- Stochastic experiment: a process in which various elementary states (or events, outcomes) are possible.
  - Toss a coin. The elementary state set ***S*** = {Head, Tail}.
  - Cast a dice. The elementary state set ***S*** = {1, 2, 3, 4, 5, 6}.
  - Text. The dictionary 



- A random variable takes on values from a set of mutually exclusive and collectively exhaustive states.

Random Variables

- Represent a variable by capital letters, e.g., X, Y, Z, ...

$$P(X), P(X, Y)$$

- Instantiations of random variables in lower case, e.g., x, y, z, ...

$$P(X = x)$$

- Random variables can be discrete or continuous.

Probability

If a variable X is discrete, $P(x)$ represents the probability of $X = x$.

$$\sum_x P(x) = 1$$

Smoking?	*Yes*	*No*
$P(x)$	0.15	0.85

Probability

Total N

$Y \backslash X$	x_1	…	x_i	…	x_Q	
y_1						
⋮						
y_j			n_{ij}			$\}\, r_j$
⋮						
y_P						
			$\underbrace{}_{c_i}$			

- **Joint Probability**

$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

- **Marginal Probability**

$$P(X = x_i) = \frac{c_i}{N}$$

$$P(Y = y_j) = \frac{r_j}{N}$$

- **Conditional Probability**

$$P(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

$$P(X = x_i | Y = y_j) = \frac{n_{ij}}{r_j}$$

Probability

Total N

$Y \backslash X$	x_1	…	x_i	…	x_Q	
y_1						
⋮						
y_j			n_{ij}			$\}\, r_j$
⋮						
y_P						
			$\underbrace{}_{c_i}$			

- **Product Rule**

$$P(X, Y) = P(Y|X)\, P(X)$$

$$P\big(X = x_i, Y = y_j\big) = \frac{n_{ij}}{N}$$

$$= \frac{n_{ij}}{c_i} \times \frac{c_i}{N} = P(Y|X)\, P(X)$$

- **Sum Rule** $P(X) = \sum_{Y} P(X, Y)$

$$P(X = x_i) = \frac{c_i}{N} = \frac{\sum_{j=1}^{P} n_{ij}}{N} = \sum_{j=1}^{P} P(X = x_i, Y = y_j)$$

Probability

- **Bayes' Theorem (Rule)**

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}$$

- **Chain Rule**

$$P(X,Y,Z,...) = P(X)P(Y \mid X)P(Z \mid X,Y)P(... \mid X,Y,Z)$$

Probability

- **Independent**

$$P(X,Y) = P(X)P(Y)$$

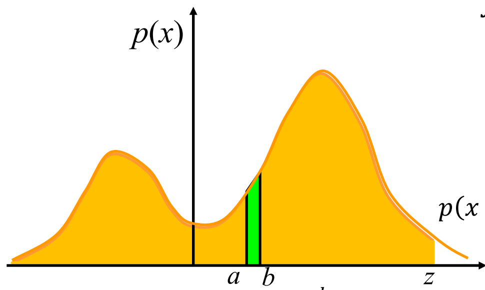
- **Conditional Independent**

$$P(X,Y \mid Z) = P(X \mid Z)P(Y \mid Z)$$

Probability

If X is continuous, $p(x) \geq 0$ is called a probability density (distribution) function (PDF).

$$\int_{-\infty}^{\infty} p(x)dx = 1$$



$$p(x \leq z) = \int_{-\infty}^{z} p(x)dx$$

$$p(a \leq x \leq b) = \int_{a}^{b} p(x)dx$$

Descriptive Statistics

Summarize a collection of data.

Expectations $\quad \mathbb{E}[f] = \sum_x p(x)f(x)$

Conditional Expectation $\quad \mathbb{E}_x[f|y] = \sum_x p(x|y)f(x)$

Approximate Expectation $\quad \mathbb{E}[f] \simeq \frac{1}{N}\sum_{n=1}^{N} f(x_n)$

Descriptive Statistics

Summarize a collection of data.

Variance: $var[f] = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2]$

Calculate *var*$[f]$ by going through the whole data set once

Covariance: $cov[x, y] = \mathbb{E}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])]$

Calculate $cov[x, y]$ by going through the whole dataset once

Vector

d-dimension vector and its transpose

$$\vec{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_d \end{bmatrix} \qquad \vec{v}^T = [v_1 \quad v_2 \quad \cdots \quad v_d]$$

Vector

$$\vec{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_d \end{bmatrix}$$

The magnitude of a vector $||\vec{v}|| = \sqrt{\sum_{k=1}^{d} (v_k \cdot v_k)}$

The inner product of two vectors

$$\vec{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_d \end{bmatrix}$$

$$\langle \vec{v}, \vec{u} \rangle = \vec{v}^T \vec{u} = \vec{u}^T \vec{v} = \sqrt{\sum_{k=1}^{d} (v_k \cdot u_k)}$$

The angle θ between two vectors satisfies:

$$cos\theta = \frac{\langle \vec{v}, \vec{u} \rangle}{||\vec{v}|| \cdot ||\vec{u}||}$$

$\vec{v}$ and $\vec{u}$ are orthogonal if $\langle \vec{v}, \vec{u} \rangle = 0$

Matrix

m×*n* matrix and its transpose

$$A_{m \times n} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \qquad (A_{m \times n})^T = \begin{bmatrix} a_{11} & \cdots & a_{m1} \\ \vdots & \ddots & \vdots \\ a_{1n} & \cdots & a_{mn} \end{bmatrix}$$

Vector & Matrix – Multiplication

The product of two matrices

$$A_{m\times n}B_{n\times q}=\begin{bmatrix}a_{11} & \cdots & a_{1n}\\ \vdots & \ddots & \vdots\\ a_{m1} & \cdots & a_{mn}\end{bmatrix}\begin{bmatrix}b_{11} & \cdots & b_{1q}\\ \vdots & \ddots & \vdots\\ b_{n1} & \cdots & b_{nq}\end{bmatrix}=\begin{bmatrix}c_{11} & \cdots & c_{1q}\\ \vdots & \ddots & \vdots\\ c_{m1} & \cdots & c_{mq}\end{bmatrix}$$

where $c_{ij}=\sum_{k=1}^{n}a_{ik}b_{kj}$

The product of matrix and vector

$$A\vec{v} \overset{?}{=} \vec{v}^{T}A$$

Big Matrix × Big Matrix

$A_{1,000,000\times 1,000,000}$ $B_{1,000,000\times 1,000,000}$

Take long time to compute.

1,000,000 × 1,000,000 = 1,000,000,000,000 ≈ 1T

Block Matrix

$$A = \left[\begin{array}{cc:cc} 1 & 1 & 2 & 2 \\ 1 & 1 & 2 & 2 \\ \hdashline 3 & 3 & 4 & 4 \\ 3 & 3 & 4 & 4 \end{array}\right]$$

$$A_{11} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \quad A_{12} = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} \quad A_{21} = \begin{bmatrix} 3 & 3 \\ 3 & 3 \end{bmatrix} \quad A_{22} = \begin{bmatrix} 4 & 4 \\ 4 & 4 \end{bmatrix}$$

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

Block Matrix Multiplication

$$A = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1s} \\ A_{21} & A_{22} & \dots & A_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ A_{q1} & A_{q2} & \dots & A_{qs} \end{bmatrix} \qquad B = \begin{bmatrix} B_{11} & B_{12} & \dots & B_{1r} \\ B_{21} & B_{22} & \dots & B_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ B_{s1} & B_{s2} & \dots & B_{sr} \end{bmatrix}$$

$$C = AB = \begin{bmatrix} C_{11} & \cdots & C_{1m} & \cdots & C_{1r} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ C_{n1} & \cdots & C_{nm} & \cdots & C_{nr} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ C_{q1} & \cdots & C_{11} & \cdots & C_{qr} \end{bmatrix} \qquad C_{nm} = \sum_{k=1}^{s} A_{nk} B_{km}$$

Machine Learning

Data-Driven Decision Making

Models

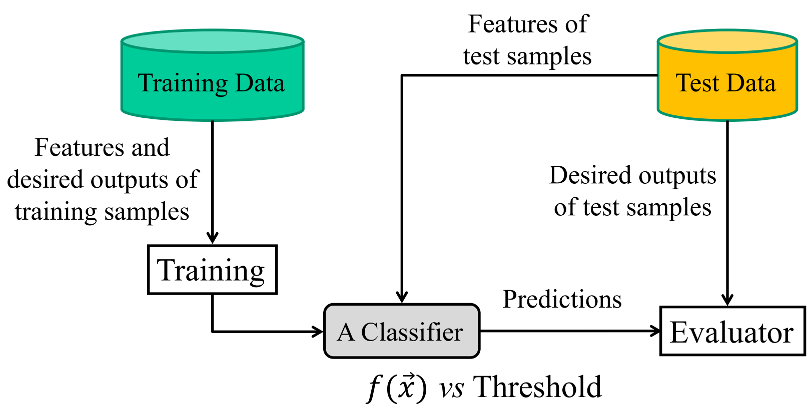
Data

Why machine learning? **Proven to be effective in practice!**

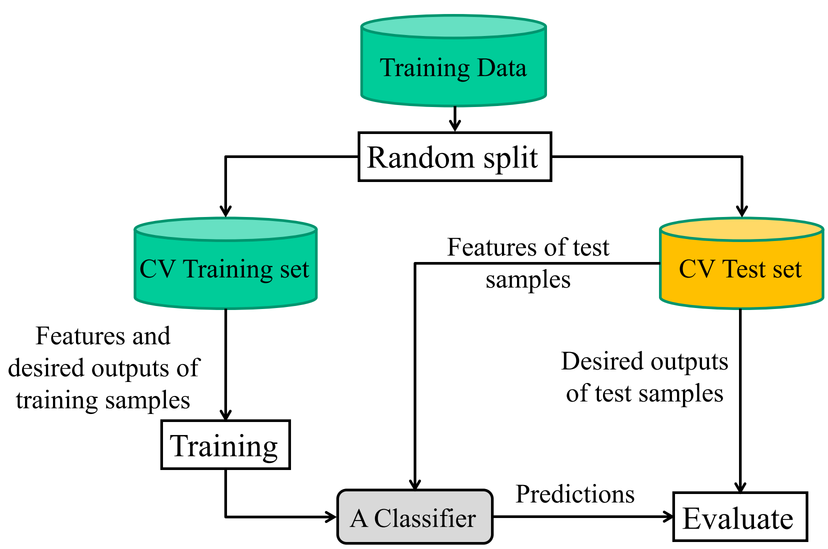
Types of Learning

- Supervised Learning: given examples of input-output pairs, train models that produce the "correct" outputs for new inputs.
- Unsupervised Learning: given only inputs as training, find structure in the training data. (e.g., clustering, probability density estimation, pattern mining, etc.).
- Reinforcement Learning: take actions in an environment so as to maximize cumulative reward.
- Mixture of the above

Evaluating Classifiers



Cross–Validation (CV)



Confusion Matrix		Ground Truth	
		P – Positive	*N* – Negative
Prediction	Positive	*TP* – True Positive	*FP* – False Positive
	Negative	*FN* – False Negative	*TN* – True Negative

$$\text{true positive rate} = \frac{TP}{P}$$

$$\text{false positive rate} = \frac{FP}{N}$$

$$\text{recall} = \frac{TP}{P}$$

$$\text{precisoin} = \frac{TP}{TP + FP}$$

$$\text{accuracy} = \frac{TP + TN}{P + N}$$

$$\text{F}-\text{measure} = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}}$$

$$\text{sensitivity} = \text{recall}$$

$$\text{specificity} = 1 - \text{false positive rate} = \frac{TN}{FP + TN}$$

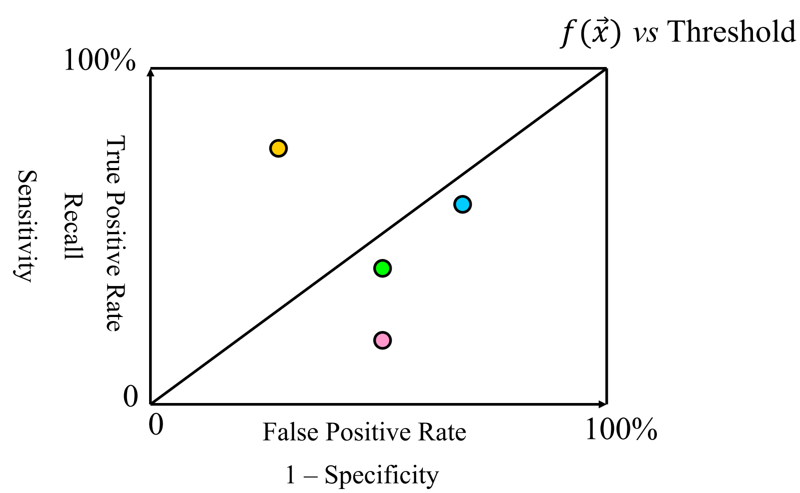
$$\text{positive predictive value} = \text{precision}$$

ROC – Receiver Operating Characteristics
A Brief History

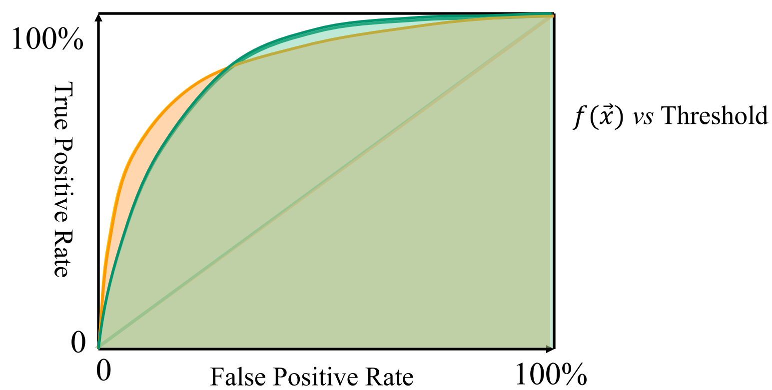
- Signal Detection: tradeoff between hit rates and false alarm rates (J. P. Egan 1975).
- Diagnostic Systems (Swets 1988).
  - Materials testing, medical imaging, weather forecasting, information retrieval, polygraph lie detection, aptitude testing
- Machine Learning: evaluate and compare algorithms (Spackman 1989).
- Medical decision making (Zou 2002)
  - http://splweb.bwh.harvard.edu:8000/pages/ppl/zou/roc.html

Yom Fawcett 2004, ROC Graphs: Notes and Practical Considerations for Researchers.

ROC



Area Under the Curve (AUC)



- AUC = the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.
- Although the ROC AUC statistic has been widely used in machine learning community for model comparison, it have also been questioned.