

L7: Probability Basics

CS 344R/393R: Robotics
Benjamin Kuipers

Outline

1. Bayes' Law
2. Probability distributions
3. Decisions under uncertainty

Probability

- For a proposition A , the probability $p(A)$ is your *degree of belief* in the truth of A .
 - By convention, $0 \leq p(A) \leq 1$.
- This is the *Bayesian* view of probability.
 - It contrasts with the view that probability is the *frequency* that A is true, over some large population of experiments.
 - The frequentist view makes it awkward to use data to estimate the value of a constant.

Probability Theory

- $p(A,B)$ is the *joint probability* of A and B .
- $p(A|B)$ is the *conditional probability* of A given B .

$$p(A|B) + p(\neg A|B) = 1$$

$$p(A,B) = p(B|A) * p(A)$$

- **Bayes Law:**

$$p(B|A) = \frac{p(A,B)}{p(A)} = \frac{p(A|B) * p(B)}{p(A)}$$

Bayes' Law for Diagnosis

- Let H be a hypothesis, E be evidence.
$$p(H|E) = \frac{p(E|H) * p(H)}{p(E)}$$
- $p(E|H)$ is the *likelihood* of the data, given the hypothesis.
- $p(H)$ is *prior* probability of hypothesis.
- $p(E)$ is prior probability of the evidence (but acts as a normalizing constant).
- $p(H|E)$ is what you really want to know (*posterior* probability of hypothesis).

Which Hypothesis To Prefer?

- Maximum Likelihood (ML)
 - $\max_H p(E|H)$
 - The model that makes the data most likely
- Maximum *a posteriori* (MAP)
 - $\max_H p(E|H) p(H)$
 - The model that is the most probable explanation
- (Story: perfect match to rare disease)

Bayes Law

- The denominator in Bayes Law acts as a normalizing constant:

$$p(H \mid E) = \frac{p(E \mid H) p(H)}{p(E)} = \eta p(E \mid H) p(H)$$

$$\eta = p(E)^{-1} = \frac{1}{\sum_H p(E \mid H) p(H)}$$

- It ensures that the probabilities sum to 1 across all the hypotheses H .

Independence

- Two random variables are *independent* if
 - $p(X, Y) = p(X) p(Y)$
 - $p(X \mid Y) = p(X)$
 - $p(Y \mid X) = p(Y)$
 - These are all equivalent.
- X and Y are *conditionally independent given Z* if
 - $p(X, Y \mid Z) = p(X \mid Z) p(Y \mid Z)$
 - $p(X \mid Y, Z) = p(X \mid Z)$
 - $p(Y \mid X, Z) = p(Y \mid Z)$
- Independence simplifies inference.

Accumulating Evidence (Naïve Bayes)

$$p(H \mid d_1, d_2, \dots, d_n) = p(H) \frac{p(d_1 \mid H) p(d_2 \mid H) \dots p(d_n \mid H)}{p(d_1) p(d_2) \dots p(d_n)}$$

$$p(H \mid d_1, d_2, \dots, d_n) = p(H) * \prod_{i=1}^n \frac{p(d_i \mid H)}{p(d_i)}$$

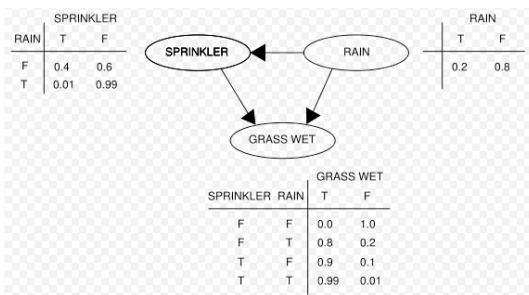
$$p(H \mid d_1, d_2, \dots, d_n) = \eta p(H) * \prod_{i=1}^n p(d_i \mid H)$$

$$\log p(H \mid d_1, d_2, \dots, d_n) = \log p(H) + \sum_{i=1}^n \log p(d_i \mid H) + \eta'$$

Bayes Nets Represent Dependence

- The nodes are random variables.
- The links represent dependence.
 - $p(X_i \mid \text{parents}(X_i))$
 - Independence can be inferred from network
- The network represents how the joint probability distribution can be decomposed.
 - $p(X_1, \dots, X_n) = \prod_{i=1}^n p(X_i \mid \text{parents}(X_i))$
- There are effective propagation algorithms.

Simple Bayes Net Example



Outline

- Bayes' Law
- Probability distributions
- Decisions under uncertainty

Expectations

- Let x be a random variable.
- The expected value $E[x]$ is the mean:

$$E[x] = \int x p(x) dx \approx \bar{x} = \frac{1}{N} \sum_1^N x_i$$
 - The probability-weighted mean of all possible values. The sample mean approaches it.
- Expected value of a vector \mathbf{x} is by component.

$$E[\mathbf{x}] = \bar{\mathbf{x}} = [\bar{x}_1, \dots, \bar{x}_n]^T$$

Variance and Covariance

- The variance is $E[(x-E[x])^2]$

$$\sigma^2 = E[(x - \bar{x})^2] = \frac{1}{N} \sum_1^N (x_i - \bar{x})^2$$
- Covariance matrix is $E[(\mathbf{x}-E[\mathbf{x}])(\mathbf{x}-E[\mathbf{x}])^T]$

$$C_{ij} = \frac{1}{N} \sum_{k=1}^N (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)$$

- Divide by $N-1$ to make the sample variance an *unbiased estimator* for the population variance.

Biased and Unbiased Estimators

- Strictly speaking, the sample variance

$$\sigma^2 = E[(x - \bar{x})^2] = \frac{1}{N} \sum_1^N (x_i - \bar{x})^2$$

is a biased estimate of the population variance. An unbiased estimator is:

$$s^2 = \frac{1}{N-1} \sum_1^N (x_i - \bar{x})^2$$

- **But:** “If the difference between N and $N-1$ ever matters to you, then you are probably up to no good anyway ...” [Press, et al]

Covariance Matrix

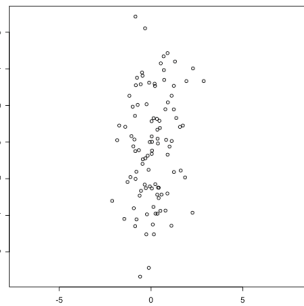
- Along the diagonal, C_{ii} are variances.
- Off-diagonal C_{ij} are essentially correlations.

$$\begin{bmatrix} C_{1,1} = \sigma_1^2 & C_{1,2} & & C_{1,N} \\ C_{2,1} & C_{2,2} = \sigma_2^2 & & \\ & & \ddots & \vdots \\ C_{N,1} & \dots & \dots & C_{N,N} = \sigma_N^2 \end{bmatrix}$$

Independent Variation

- x and y are Gaussian random variables ($N=100$)
- Generated with $\sigma_x=1$ $\sigma_y=3$
- Covariance matrix:

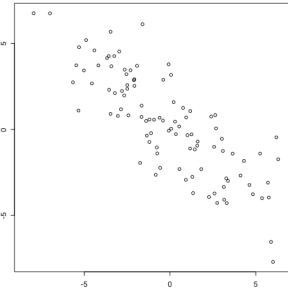
$$C_{xy} = \begin{bmatrix} 0.90 & 0.44 \\ 0.44 & 8.82 \end{bmatrix}$$



Dependent Variation

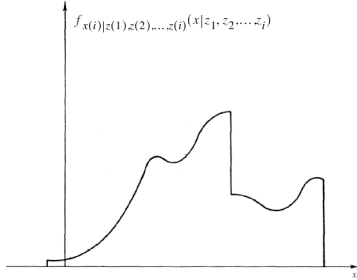
- c and d are random variables.
- Generated with $c=x+y$ $d=x-y$
- Covariance matrix:

$$C_{cd} = \begin{bmatrix} 10.62 & -7.93 \\ -7.93 & 8.84 \end{bmatrix}$$



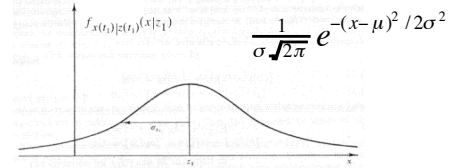
Estimates and Uncertainty

- Conditional probability density function



Gaussian (Normal) Distribution

- Completely described by $N(\mu, \sigma)$
 - Mean μ
 - Standard deviation σ , variance σ^2

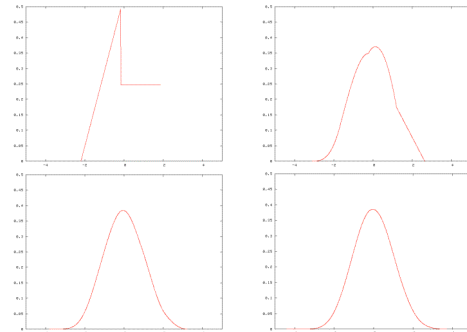


The Central Limit Theorem

- The sum of many random variables
 - with the same mean, but
 - with arbitrary conditional density functions,
 converges to a Gaussian density function.
- If a model omits many small unmodeled effects, then the resulting error should converge to a Gaussian density function.

Illustrating the Central Limit Thm

- Add 1, 2, 3, 4 variables from the same distribution.



Detecting Modeling Error

- Every model is incomplete.
 - If the omitted factors are all small, the resulting errors should add up to a Gaussian.
- If the error between a model and the data is not Gaussian,
 - Then some omitted factor is not small.
 - One should find the dominant source of error and add it to the model.

Outline

1. Bayes' Law
2. Probability distributions
- 3. Decisions under uncertainty**

Diagnostic Errors and Sensor Interpretation

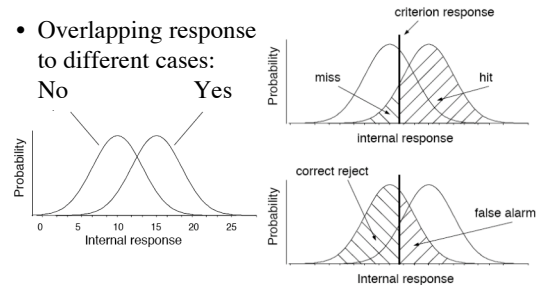
- Interpreting sensor values is like diagnosis.

	Test=Pos	Test=Neg		
Disease present	True Positive	False Negative	hit	miss
Disease absent	False Positive	True Negative	false alarm	correct reject

- Every test has false positives and negatives.
 - Sonar(fwd)= d implies Obstacle-at-distance(d) ??

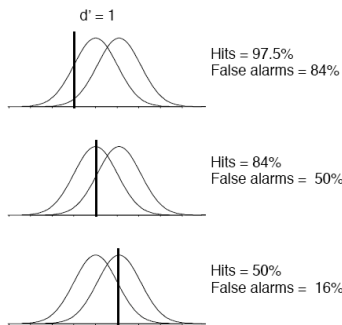
Tests: Sensor Noise and Decision Thresholds

- Overlapping response to different cases:



The Test Threshold Requires a Trade-Off

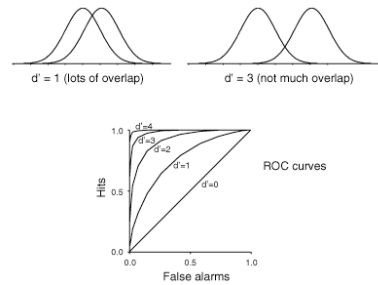
- You can't eliminate all error.
- Choose which errors are important



ROC Curves

$$d' = \frac{\text{separation}}{\text{spread}}$$

- The overlap d' controls the trade-off between types of errors.



- For more, search on *Signal Detection Theory*.

Bayesian Reasoning

- One strength of Bayesian methods is that they reason with probability distributions, not just the most likely individual case.
- For more, see Andrew Moore's tutorial slides
 - <http://www.autonlab.org/tutorials/>
- Coming up:
 - Regression to find models from data
 - Kalman filters to track dynamical systems
 - Visual object trackers.