Short Probability Overview

Goals

Review the basics of probability distributions and statistics

Probability overview

Random variables, discrete probabilities, probability densities, cumulative distributions

Discrete: random variable X can take on a finite set of discrete values

X = {x(1),...,x(N)]

$$\sum_{i=1}^{N} p_i = \sum_{i=1}^{N} p(X = x(i)) = 1$$

Densities: X takes on continuous values, x, in some range.

Density : $p(x)$

Analogous to material mass,
we can think of the probability over some small domain of the random variable as " probability mass " :

$$\text{prob}(x < X < \text{dx} + x) = \int_{X}^{\text{dX}+X} p(x)\, dx$$

$$\text{prob}(x < X < \text{dx} + x) \simeq p(x)\, \text{dx}$$

With the mass analogy, however, an object (event space) always " weighs " 1 :

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

Cumulative distribution:

$$\text{prob}\,(X < x) = \int_{-\infty}^{x} p(X)\, dX$$

Densities of discrete random variables

The Dirac Delta function, $\delta[\bullet]$, allows us to use the mathematics of continuous distributions for discrete ones, by defining the density as:

$$p[x] = \sum_{i=1}^{N} p_i \delta[x - x[i]], \text{ where } \delta[x - x[i]] = \begin{cases} \infty \text{ for } x = x[i] \\ 0 \text{ for } x \neq x[i] \end{cases}$$

Think of the delta function, $\delta[\bullet]$, as ϵ wide and $1/\epsilon$ tall, and then let ϵ -> 0, so that:

$$\int_{-\infty}^{\infty} \delta(y)\, dy = 1$$

The density, p[x], is a series of spikes. It is infinitely high only at those points for which x = x[i], and zero elsewhere. But "infinity" is scaled so that the local mass or area around each spike x[i], is p_i.

Joint probabilities

Prob (X AND Y) = $p(X, Y)$
Joint density : $p(x, y)$

Three basic rules of probability

Suppose we know everything there is to know about a set of variables (A,B,C,D,E). What does this mean in terms of probability? It means that we know the joint distribution, p(A,B,C,D,E). In other words, for any particular combination of values (A=a,B=b, C=c, D=d,E=e), we can calculate, look up in a table, or determine some way or another the number p(A=a,B=b, C=c, D=d,E=e).
Deterministic relationships are special cases.

Rule 1: Conditional probabilities from joints: The product rule

Probability about an event changes when new information is gained.
Prob(X given Y) = p(X|Y)

$$p(X \mid Y) = \frac{p(X, Y)}{p(Y)}$$

$$p(X, Y) = p(X \mid Y)\, p(Y)$$

The form of the product rule is the same for densities as for probabilities.

Rule 2: Lower dimensional probabilities from joints: The sum rule (marginalization)

$$p(X) = \sum_{i=1}^{N} p(X, Y(i))$$

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

Rule 3: Bayes' rule

From the product rule, and since p[X,Y] = p[Y,X], we have:

$$p(Y \mid X) = \frac{p(X \mid Y)\, p(Y)}{p(X)}, \text{ and using the sum rule, } p(Y \mid X) = \frac{p(X \mid Y)\, p(Y)}{\sum_Y p(X, Y)}$$

Bayes Terminology in inference

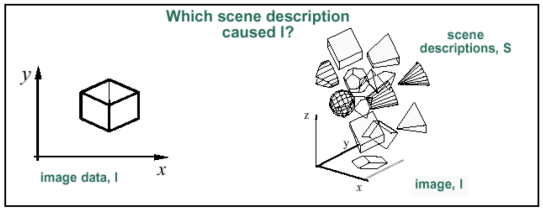
Suppose we have some partial data (see half of someone's face), and we want to recall or complete the whole. Or suppose that we hear a voice, and from that visualize the face. These are both problems of statistical inference. We've already studied how to complete a partial pattern using energy minimization,

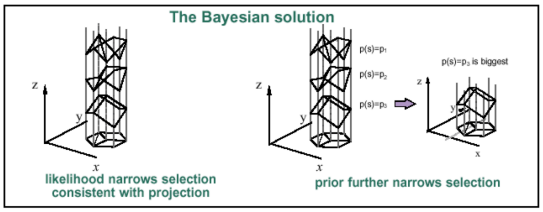
and how energy minimization can be viewed as probability maximization.
We typically think of the **Y** term as a random variable over the hypothesis space (a face), and **X** as data or a stimulus (partial face, or sound). So for recalling a pattern **Y** from an input stimulus **X,** We'd like to have a function that tells us:
p(Y | X) which is called the **posterior** probability of the hypothesis (face) given the stimulus (partial face or sound).
-- i.e. what you get when you condition the joint by the stimulus data. The posterior is often what we'd like to base our decisions on, because it can be proved that picking the hypothesis **Y** which maximizes the posterior (i.e. maximum a posteriori or **MAP** estimation) minimizes the average probability of error.
p(Y) is the **prior** probability of the hypothesis. Some hypotheses are more likely than others. Given a context, such as your room, some faces are a priori more likely than others. For me an image patch stimulating my retina in my kitchen is much more likely to be my wife's than my brother's (who lives in another state). This shows that priors are contingent, i.e. conditional on context, **p(Y| context)**.
p(X|Y) is the **likelihood** of the hypothesis. Note this is a probability of **X**, but not of **Y**.(The sum over X is one, but the sum over Y isn't necessarily one.)

Bayes Terminology in visual perception

$$p[S \mid I] = \frac{p[I \mid S]\, p[S]}{p[I]}$$

Usually, we will be thinking of the **Y** term as a random variable over the hypothesis space, and **X** as data. So for visual inference, **Y = S** (the scene), and **X = I** (the image data), and **I = f(S)**.
We'd like to have:
p(S|I) is the **posterior** probability of the scene given the image
-- i.e. what you get when you condition the joint by the image data. The posterior is often what we'd like to base our decisions on, because as we discuss below, picking the hypothesis **S** which maximizes the posterior (i.e. maximum a posteriori or **MAP** estimation) minimizes the average probability of error.
p(S) is the **prior** probability of the scene.
p(I|S) is the **likelihood** of the scene. Note this is a probability of **I**, but not of **S**.





Independence

Knowledge of one event doesn't change the probability of another event.
p(X)=p(X|Y)

p(X,Y)=p(X)p(Y)

Density mapping theorem

Suppose we have a change of variables that maps a discrete set of x's uniquely to y's: X->Y.

Discrete random variables

No change to probability function. The mapping just corresponds to a change of labels, so the probabilities p(X)=p(Y).

Continuous random variables

Form of probability density function does change because we require the probability "mass" to be unchanged: p(x)dx = p(y)dy

Suppose, y=f(x)

$$p_Y(y)\,\delta y = p_X(x)\,\delta x$$

In higher dimensions, the transformation is done by multiplying the density by the Jacobian, the determinant of the matrix of partial derivatives of the change of coordinates.

$$p_Y(y)=\int \delta(y-f(x))\,f^{-1}(x)\,p_X(x)\,dx$$

over each monotonic part of f.

Convolution theorem for adding rvs

Let x be distributed as g(x), and y as h(x). Then the probability density for z=x+y is, f(z):

$$f(z) = \int g(s)\,h(z-s)\,ds \tag{1}$$

Statistics

Expectation & variance

Analogous to center of mass:

Definition of expectation or average:

$$\text{Average}[X] = \overline{X} = E[X] = \Sigma\, x[i]\, p[x[i]] \sim \sum_{i=1}^{N} x_i / N$$

$$\mu = E[X] = \int x\,p(x)\,dx$$

Some rules:

E[X+Y]=E[X]+E[Y]
E[aX]=aE[X]
E[X+a]=a+E[X]

Definition of variance:

σ^2 = Var[X] = E[[X-μ]^2] = $\sum_{j=1}^{N} p(x(j)\,(x(j) - \mu)^2 = \sum_{j=1}^{N} p_j(x_j - \mu))^2$

$$\text{Var}[X] = \int (x - \mu)^2\, p(x)\, dx \sim \sum_{i=1}^{N} (x_i - \mu)^2 / N$$

Standard deviation:

$$\sigma = \sqrt{\text{Var}[X]}$$

Some rules:

$$\text{Var}[X] = E\left[X^2\right] - E[X]^2$$

$$\text{Var}[aX] = a^2\, \text{Var}[X]$$

Covariance & Correlation

Covariance:

Cov[X,Y] =E[[X - μ_X] [Y - μ_Y]]

Correlation coefficient:

$$\rho[X, Y] = \frac{\text{Cov}[X, Y]}{\sigma_X\, \sigma_Y}$$

Cross and Autocovariance matrix

Suppose X and Y are vectors: {X_1, X_2,...} and {Y_1, Y_2, ...}

Cov[X_i,Y_j] =E[[X_i - μ_{X_i}] [Y_j - μ_{Y_j}]] ~ $\sum_{n=1}^{N} (x_i^n - \mu_{X_i})\,(y_j^n - \mu_{Y_j})^T / N$

Autocov[X_i,X_j] =E[[X_i - μ_{X_i}] [X_j - μ_{Y_j}]] ~ $\sum_{n=1}^{N} (x_i^n - \mu_{X_i})\,(x_j^n - \mu_{X_i})^T / N$

In other words, the autocovariance matrix can be approximated by the average outer product. It is a Hebbian matrix memory of pair-wise relationships.

Independent random variables

If p(X,Y)=p(X)p(Y), then

$$E[X\,Y] = E[X]\, E[Y] \text{ (uncorrelated)}$$

$$\text{Cov}[X, Y] = \rho[X, Y] = 0$$

$$\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y]$$

If two random variables are uncorrelated, they are not necessarily independent.
Two random variables are said to be orthogonal if their correlation is zero.

Degree of belief vs., relative frequency

What is the probability that the Vikings will win the Superbowl in 2004? Assigning a number between 0 and 1 is assigning a degree of belief. These probabilities are also called subjective probabilities. "Odds" determine subjective probabilities, where the "odds of **x** to **y**" means probability = **x/(x+y)**.
What is the probability that a coin will come up heads? In this case, we can do an experiment. Flip the coin n times, and count the number of heads, say h[n], and then set the probability, p = h[n]/n -- the relative frequency . Of course, if we did it again, we may not get the same estimate of p. One solution

often given is:

$$p = \lim_{n\to\infty} \frac{h(n)}{n}$$

A problem with this, is that there is no guarantee that *a* well – defined limit exists.

In some domains we can measure statistics, and model probabilities of both inputs and outputs. So the relative frequency interpretation seems reasonable. In practice, the dimensions of many problems in perception, cognition, and memory are so high, that it is impractical to do this. Once we use the statistical framework to model perception, say of a particular cue (say), then probabilities are more like "subjective unconscious beliefs".

Principle of insufficient reason

Principle of symmetry

Suppose we have N events, x[1],x[2],x[3],...,x[N] that are all physically identical except for the label. Then assume that

$$\text{prob}(x(1)) = \text{prob}(x(2)) = \text{prob}(x(3)) = \text{prob}(x(N)) = \frac{1}{N}$$

In other words,if we have no additional information about the events,we should assume that they are uniformly distributed. I.e., assume a uniform prior.
What about the continous case where there is no reason to assume any particular value at all between -∞ and +∞?
Improper priors.

Information theory and Maximum entropy

Information theory provides a powerful extension to the principle of symmetry. Information of event X is:

$$\text{Information}[X] = -\log_2(p(X))$$

Using the definition of expectation above, we can specify the expectation of information, which is called entropy. Entropy of a random variable X with probability distribution p[X] is:

$$H(X) = \text{Average}(\text{Information}[X]) = -\sum_X p(X)\log_2(p(X))$$

It can be shown that out of all possible probability distributions, H(X) is biggest for the uniform distribution, p(X)=1/N. Maximum entropy is looking like symmetry.
It turns out that a more powerful formulation of the principle of symmetry is maximum entropy. For example, out of all possible probability distributions of a random variable with infinity range, but with a specific mean and standard deviation, the Gaussian is unique in having the largest entropy. If the range goes from zero to infinity, and we know the mean, the maximum entropy distribution is an exponential (Cover and Thomas).

An interesting application of the maximum entropy principle is to learning image textures joint probabilities: p(I[1],...,I[N]), where N is very big, but where one has only a relatively small number of measured statistics relative to the number of possible images (which is really huge). The measurements underdetermine the dimensionality of the probability space--i.e. there are many different probability distributions which give the same statistics. So the principle of symmetry, or insufficient reason, says to choose the

one with the maximum entropy.

References

Applebaum, D. (1996). Probability and Information . Cambridge, UK: Cambridge University Press.
Cover, T. M., & Joy, A. T. (1991). *Elements of Information Theory*. New York: John Wiley & Sons, Inc.
Duda, R. O., & Hart, P. E. (1973). Pattern classification and scene analysis . New York.: John Wiley & Sons.
Kersten, D. and P.W. Schrater (2000), *Pattern Inference Theory: A Probabilistic Approach to Vision*, in *Perception and the Physical World*, R. Mausfeld and D. Heyer, Editors. , John Wiley & Sons, Ltd.: Chichester. (pdf)
Kersten, D., & Yuille, A. (2003). Bayesian models of object perception. Current Opinion in Neurobiology, 13(2) http://gandalf.psych.umn.edu/~kersten/kersten-lab/papers/KerstenYuilleApr2003.pdf
Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge, UK: Cambridge University Press.
Van Trees, H. L. (1968). Detection, Estimation and Modulation Theory . New York: John Wiley and Sons.