Course notes, CS664, 11/18/04

- Quiz 4 today at end of class

- MRI guest lecture on Tuesday (RDZ is out of town).

- Next Tuesday (11/30): Quiz 5, also hand in 1-page paper summary via CMS. Graded by Thursday.

Random variables

- Consider an RV X. In the discrete world, this assigns an integer to every outcome. The event $X = 3$ is a set of outcomes, with a probability. We can change 3 to 4 to 5, and plot this as a function (call its argument α). It's a PMF (Probability Mass Function) $f_X(\alpha)$.

- You can also consider the events $X \leq 3$, $X \leq 4$ etc, and plot their probabilities. This gives you a (cumulative) distribution $F_X(\alpha)$. Discretely,

$$F_X(\alpha) = \sum_{\beta=-\infty}^{\alpha} f_X(\alpha).$$

- The important thing about an RV is the probability that it lies in some interval. The probability that $3 \leq X \leq 4$ is

$$f_X(3) + f_X(4) = F_X(4) - F_X(2)$$

- There are continuous versions of these identities for real-valued RV's. Note that you can no longer think of $f_X(\alpha)$ as the probability that X takes on the value α — for a real-valued variable this is always 0. Instead you have to think of it as defining the probability in an interval

$$Pr(a \leq X \leq b) = \int_{\alpha=a}^{\alpha=b} f_X(\alpha) d\alpha.$$

Density estimation

- Let's suppose that we have a RV X with some density f_X. Imagine that we draw samples according to this density. Easy in the discrete case — just generate a random number z uniformly between 0 and 1, and the sample value will be the unique (why?) α such that $F_X(\alpha) = z$. Similar for continuous case, but there are always numerical errors...

- Now imagine that we are given a bunch of data points sampled from an unknown density. How do we compute the density?

[Note: we will assume throughout that the samples are generated independently.]

- This is a simple example of statistical estimation, a problem that shows up again and again in vision (and elsewhere!)

- Simplest vision example: a "color" is a PDF in (say) RGB space, and you're given a (1-color) image. Each pixel is drawn from the PDF. How do you compute the PDF?

Parametric density estimation

- Sometimes you know the PDF's form, just not its parameters (arguments). Classical example: the PDF might be Gaussian with unknown mean μ and variance σ^2. Or it might be uniform in the interval $[a, b]$ with these 2 parameters.

- In a given run, we might see a roughly bell-shaped set of data centered around μ with width σ^2. Or we might see about the same amount of data in $[a, b]$ and nothing elsewhere.

- How do we solve for the parameters? The standard solution is quite general (again, due to Gauss).

- For a particular hypothesized distribution μ, σ^2 you can compute the probability that it would generate the data that you saw. This *very important* function is called the likelihood, and even has its own Latex macro $\ell(\mu, \sigma^2) = Pr(data|\mu, \sigma^2)$.

- The likelihood, as we will see, more or less encapsulates a model of measurement error.

- The obvious idea is that the right set of parameters made the observations relatively likely. For any infinite-tailed distribution, *any* data is possible, just with very low probability. The Gaussian PDF centered at zero can generate uniform data centered at -100, it's just not very likely.

- Maximum likelihood estimation: the best set of parameters maximizes ℓ. Note that the justification is philosophical, not mathematical (i.e., non-rigorous). But it's actually quite intuitive.

- It is not hard to prove that the mean of the data

$$\bar{d} = \frac{\sum_i d_i}{N}$$

is the ML estimate of μ, and that the variance of the data

$$\frac{\sum_i (d_i - \bar{d})^2}{N-1}$$

is the ML estimate of σ^2.

- This is where the ambiguity about N versus $N-1$ in the definition of the std deviation arises. Using N gives a slightly too small ("biased") estimate of σ^2.

Maximum likelihood estimation

- We can think of an experimental result as an outcome, with a (hidden) true state and an observed state. Think of observing a single pixel intensity (image denoising), where there was also a true intensity. An outcome is $(true, observed)$.

- The likelihood is $\ell(true) = Pr(observed|true)$, which is basically a noise model. For instance with Gaussian noise we have

$$\ell(I_t) \propto \exp\Big(-\frac{(I_t - I_o)^2}{2\sigma^2}\Big)$$

- Of all the hypotheses about the true intensity I_t, ML suggests we pick the one that maximizes $\ell(I_t)$.

- An alternative way to maximize the likelihood is to maximize the log-likelihood, hence to minimize

$$\frac{(I_t - I_o)^2}{2\sigma^2}$$

  which happens at input equals output.

- You can think of the likelihood as a probability distribution in a row (or column) of the outcomes. Called a *marginal*.

- ML is very powerful. For instance, line fitting using least squares is equivalent to ML estimation of the line parameters, assuming gaussian noise. The larger the residual at a point the lower the likelihood that it could have been generated by this line.

- Aside: why does everone love gaussian noise? Central limit theorem!

- There is a variant called M-estimation which is ML-estimation under the contamination model. And an even better varient called Least Median Squares (think of this as fitting the thinnest ruler to cover half the data).

Non-parametric density estimation

- Back to density estimation. In vision, densities are often highly non-Gaussian. As a rule we don't have a model for them. Moreover, they often have multiple modes (competing hypotheses).

- What to do? Obvious solution is to take the data and "blur" it to created a PDF. Non-parametric density estimation, aka kernel methods, aka Parzen estimation.

- Note that this is necessary even in the discrete case.

- Why do you want to estimate the density? Often in practice you'd like to find the modes of the density. These might well be colors. Note that this is highly related to clustering (in fact, one

way to do clustering is with statistical estimation methods).

Mean shift

- The mean shift is a way to compute the mode without estimating the density. Assumption: the density of the data increases as you move towards the mode. This is often true, but not always...

- You take a window and compute the difference between the center of the window and the center of the data. This gives you a shift. You can prove that this points in the direction of the gradient (increasing density of the PDF).

- A very simple idea with lots of nice applications.