Bayes decision theory and rational decisions

- Previous lectures gave examples where linear, or non-linear, filtering was followed by a decision process. Examples included binocular stereo, motion estimation, and regression.
- Bayes decision theory gives a framework for making rational/optimal decisions in the presence of uncertain information. It was developed in the second world was for applications such as interpreting radar signals and decrypting codes.
- It has been proposed as a theory for how humans make rational decisions, though experiments by Tversky and Kahneman suggest it is not the whole story.
- Bayes decision theory related closely to other disciplines like Signal detection theory (Psychology) and Machine Learning. But it has limitations which will be discussed later.

Bayes decision theory and ideal observers

- Bayes decision theory is a framework for making optimal decisions in the presence of uncertainty. We represent the input by $x \in \mathcal{X}$ and the output by $y \in \mathcal{Y}$ (e.g., for edge detection x is the filter response $f(I)$, and $y \in \{\pm 1\}$ indicates if an edge is present or not).
- We assume that there is a probability distribution $P(x, y)$ that generates the input and output. This can be expressed in terms of a *prior* $P(y)$ and a *likelihood* $P(x|y)$ by the identity $P(x, y) = P(x|y)P(y)$. A decision rule is expressed as $\hat{y} = \alpha(x)$. We specify a *loss function* $L(\alpha(x); y)$, which is the cost of making decision $\alpha(x)$ if the real decision should be y.
- The *risk* is specified by $R(\alpha) = \sum_{x,y} P(x, y)L(\alpha(x), y)$. The *Bayes rule* is $\hat{\alpha} = \arg\min_{\alpha} R(\alpha)$. The *Bayes risk* is $\min_{\alpha} R(\alpha) = R(\hat{\alpha})$.

Bayes rule (I)

- The Bayes rule is the best decision rule you can make (*subject to this criterion*) and the Bayes risk is the best performance. Hence Bayes decision theory can specify the optimal way to estimate y from input x.
- There are several important special cases. If the loss function penalizes all errors by the same amount, i.e., $L(\alpha(x), y) = K_1$ if $\alpha(x) \neq y$ and $L(\alpha(x), y) = K_2$ if $\alpha(x) = y$ (with $K_1 > K_2$), then the Bayes rule corresponds to the *maximum a posteriori* estimator $\alpha(x) = \arg\max P(y|x)$, where $P(y|x) = \frac{P(x|y)P(y)}{P(x)}$ is the *posterior* distribution of y conditioned on y.
- If, in addition, the prior is a uniform distribution, i.e., $P(y) = \mathit{constant}$, then Bayes rule reduces to the *maximum likelihood* estimate $\alpha(x) = \arg\max P(x|y)$.

Bayes rule (II)

- For binary decision problems $y \in \{\pm 1\}$, the loss function is usually chosen to pay no penalty if the correct decision is made (i.e., $\alpha(x) = y$ -1) but has a penalty F_p for *false positives*, where $y = -1$ but $\alpha(x) = 1$, and F_n for *false negatives*, where $y = 1$ but $\alpha(x) = -$ (it is assumed here that the *target* is $y = 1$ and the *distracter* is $y = -1$, so a false positive occurs if we decide that a distracter is a target, and a false negative if we decide that a target is a distracter).
- It follows that we can express the Bayes rule in terms of a log-likelihood ratio test $\log \frac{P(x|y=1)}{P(x|y=-1)} > T$, where T depends on the prior $p(y)$ and the loss function $L(\alpha(x), y)$.

Bayes rule (III)

- More specifically, the Bayes risk is $R(\alpha) = \sum_x p(x) \sum_y L(\alpha(x), y) p(Y|x)$. Then we divide the data (x, y) into four sets: (1) the *true positives* $\{(x, y) : \text{ s.t. } \alpha(x) = y = 1\}$; (2) the *true negatives* $\{(x, y) : \text{ s.t. } \alpha(x) = y = -1\}$; (3) the *false positives* $\{(x, y) : \text{ s.t. } \alpha(x) = 1, y = -1\}$; and (4) the *false negatives* $\{(x, y) : \text{ s.t. } \alpha(x) == -1, y = 1\}$. These four cases correspond to loss function values $L(\alpha(x) = 1, y = 1) = T_p$, $L(\alpha(x) = -1, y = -1) = T_n$, $L(\alpha(x) = 1, y = -1) = F_p$, $L(\alpha(x) = -1, y = 1) = F_n$ respectively. Then the decision rule $\alpha_T(.)$ reduces to:

$$\log \frac{P(x|y=1)}{P(x|y=-1)} > \log \frac{T_n - F_p}{T_p - F_n} + \log \frac{P(y=-1)}{P(y=1)}.$$

- The intuition is that the evidence in the log-likelihood must be bigger than our prior biases while taking into account the penalties paid for different types of mistakes.

Bayes rule (IV)

The results in the previous section on edge detection and texture classification can be derived from decision theory. The priors $P(y)$ specify the probability that an image patch contains an edge (empirically $P(y=1) \approx 0.05$ and $P(y=-1) \approx 0.95$). The loss function should be chosen to specify the cost of making different types of mistakes. For texture classification, the variable y takes values in a set $\mathcal{Y}$, which is called a multiclass decision. The same theory applies to tasks for which we need to make a set of related but nonlocal decisions.

Signal detection theory (I)

We now show that an important special case of *signal detection theory* (Green & Swets, 1966) – often used as a framework to model how humans make decisions when performing visual, auditory, and other tasks – can be obtained as a special case of Bayes decision theory. We consider the two class case, where $y \in \{\pm 1\}$, and suppose that the likelihood functions are specified by Gaussian distributions, $P(x|y) = \frac{1}{\sqrt{2\pi}\sigma_y} \exp\{-(x-\mu_y)^2/(2\sigma_y^2)\}$, which differ by their means (μ_1, μ_{-1}) and their variances $(\sigma_1^2, \sigma_{-1}^2)$. The Bayes rule can be expressed in terms of the log-likelihood ratio test:

$$\hat{\alpha}(x) = \arg\max_y y\{-(x-\mu_1)^2/(2\sigma_1^2) - \log\sigma_1 + (x-\mu_{-1})^2/(2\sigma_{-1}^2) + \log\sigma_2 - T\}.$$

Signal detection theory (II)

- This decision rule requires determining whether the data point x is above or below a quadratic polynomial curve in x. In the special case when the standard deviations are identical $\sigma_1^2 = \sigma_2^2$ (so we drop the subscripts $_{1,-1}$), the decision is based only on whether the data point x satisfies:

$$2x(\mu_1 - \mu_{-1}) + (\mu_1^2 - \mu_2^2) < 2T\sigma^2$$

- This special case, with $\sigma_1^2 = \sigma_{-1}^2$, is much studied in signal detection theory (Green & Swets, 1966). It means that the decision is based on a single function $d' = \frac{\mu_1 - \mu_{-1}}{\sigma}$. This quantity is used to quantify human performance for psychophysical tasks.

Ideal observer (I)

- This motivates the idea of an *ideal observer*. An observer like this has optimal performance which requires exploiting the statistical properties of the distribution $P(x, y)$ of the data.
- A classic example of ideal observer theory shows that under certain conditions, photoreceptors in the retina are almost *optimal* at detecting the photons that reach them (Barlow, 1962; Pelli, 1990). This takes into account the probability of the photoreceptors *firing* x if it receives a photon, $P(x|y = 1)$, and the probability that the photoreceptor fires spontaneously, $P(x|y = -1)$.

Ideal observer (II)

- Ideal observers can also be defined for other vision tasks (Tjan et al., 1995; Gold et al,. 2012; Trenti et al., 2010; Geisler, 2011). The difficulty, however, is judging whether humans are adapted to doing the task. It is possible to define ideal observers when human performance is much worse than the ideal observers (Watson et al., 1983).
- Why can this happen? The task may provide information for which humans are not adapted (e.g., visual inspection of circuit boards to find deficits). Also, the ideal observers know the distributions $P(x, y)$ that, for synthetic stimuli, are those chosen by the scientist performing the experiment and may have little similarity to the natural statistics of stimuli of the world, which human vision has probably adapted to.
- Barlow and Tripathy (1997) developed an ideal observer model for motion correspondence. Analysis by Lu and Yuille (2006) showed that human perception for this task was much more consistent with a model that makes generic assumption about motion (e.g., motion tends to be slow and spatially smooth).

Receiver operating characteristic curve

- Another important concept is the receiver operating characteristic (ROC) curve. This allows us to study decisions when we do not want to restrict ourselves to specific priors and loss functions. Instead, we plot the *true positive rate* as a function of the *false positive rate* by allowing the decision threshold T to vary. For each value T of the threshold, we have a decision rule $\alpha_T(.)$, which results in a fraction of *true positives* $\sum_{x:\alpha_T(x)=1} P(x|y=1)$ and *false positives* $\sum_{x:\alpha_T(x)=1} P(x|y=-1)$. This gives a single point on the ROC curve. We plot the curve by allowing T to vary. Observe that for very large T (as $T \mapsto \infty$), the true positive and false positive rates will tend to 0. While as T gets very small ($T \mapsto -\infty$), both rates will tend to 1. Hence the ROC illustrates the trade-off between the two rates.
- Bayes decision theory can be extended in a straightforward manner if the output y takes multiple values. In particular, it applies when we have a set of decision variables defined on each lattice site of an image.

When should Bayes Decision Theory be Used?

- Bayes Decision Theory (BDT) says you seem penalize the average loss. But in some situations it may be better to penalize the worst case loss. You do not want your automated car to drive over a baby sitting in the road, even though the expected risk of this is infinitesimally small. If you are making decisions against an adversary then you need to take into account their decision strategy also, this leads to game theory.
- The average loss may be sensible if you are making the same decision repeatedly, which is the case for many visual tasks, but it may not be appropriate if you make the decision only a few times (e.g., buying a house).
- How do you know the probability distributions? In practice, you have access to observations and have to estimate the probability distributions from them. Many works on machine learning involve learning the decision rule directly from data without estimating probabilities.
- Can you compute the best decision? This is easy for the simple examples we have discussed. But BDT can be applied to much more complicated situations (see later in the course) and computing the best decision may be impossible and, at best, we can compute approximations.

Bayes Risk and Empirical Risk: Machine Learning

- Bayes Decision Theory assumes that you know the probability distributions $P(x, y)$. But in practice you will only have a set of training samples $\mathcal{X} = \{(x_1, y_1), ..., (x_n, y_n)\}$. We can define the empirical loss, which is the expected loss of a decision rule $y = \alpha(x)$ with respect to the training sample, given by $1overn \sum_{i=1}^{n} L(y_i, x_i)$.
- Some forms of Machine Learning, e.g., Support Vector Machines, argue that you should determine the decision rule by minimizing the empirical loss with respect to $\alpha(.)$, Often a regularizer $R(\alpha)$ is added to the loss in order to regularize the loss function.
- One motivation for this is that you may only have a limited number n of training samples so you should try to maximize their use by directly estimating the decision rule, instead of possibly wasting samples by first estimating the probability distributions and then deducing the decision rule. An attractive aspect of support vector machines is that the decision rule depends only on data points which lie on the decision boundary. This is described in the optional reading.
- Theoretical results (PAC theory) put upper bounds on the numbers of training samples you need in order to learn a decision rule which, with high probability will perform well on new and unseen data from the same underlying data source (this assumes that there is some unknown distribution $P(x, y)$ and the training and testing data are random independent samples from this distribution. The optional reading shows how these types of theoretical results can be obtained.