Bayesian Decision Theory.

Probability theory is the framework for making decisions under uncertainty.
For classification, Bayes' rule is used to calculate the probabilities of classes.
More generally, make decisions by minimizing the expected risk.

Assume data is generated by a random process — or by a deterministic process that we only partially know.

\( X \) is a random variable
\[ X = \begin{cases} 1 & \text{heads} \\ 0 & \text{tails} \end{cases} \]
\[ \Pr(X=1) = p_0 \quad \text{and} \quad \Pr(X=0) = 1 - \Pr(X=1) = 1 - p_0. \]

If we know \( \Pr(X) \), then we can predict:
\[ \text{If } p_0 > \frac{1}{2}, \text{ predict heads} \]
\[ \text{otherwise, predict tails} \]

If we do not know \( \Pr(X) \), then learn it from training samples:
\[ \hat{p}_0 = \frac{\text{# tosses which are heads}}{\text{# tosses}} \]
Classification.

Credit scoring - is the customer a credible risk?

\[ C = 1 \] high-risk customer
\[ C = 0 \] low-risk customer.

\[ X = (x_1, x_2) \]
\[ x_1 \] - income
\[ x_2 \] - savings

Suppose we know \( P(C | x_1, x_2) \)
and get a new customer with \((x_1, x_2)\)

choose \( \begin{cases} C = 1, & \text{if } P(C=1 | x_1, x_2) > 0.5 \\ C = 0, & \text{otherwise} \end{cases} \)

Equivalently
\[ \begin{cases} C = 1, & \text{if } P(C=1 | x_1, x_2) > P(C=0 | x_1, x_2) \\ C = 0, & \text{otherwise} \end{cases} \]

Prob of error = \[ 1 - \max \{ P(C=1 | x_1, x_2), P(C=0 | x_1, x_2) \} \]

How to calculate \( P(C|X) \)? \( X = (x_1, x_2) \)

Bayes Rule
\[ P(C|X) = \frac{P(C)P(X|C)}{P(X)} \]

\( P(C) \) - prior probability (before observing data)
\( P(x|c) \) - class likelihood.
\( P(x) \) - evidence
(3) 

\[
\text{Classifier (cont)}
\]

\[
\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}
\]

Now, we assume we know the prior and likelihood (later in the course, we will learn it).

In general case, with \( K \) classes \( C_i : i = 1 \ldots K \)

\[
\text{prior, } P(C_i) > 0, \quad \sum_{i=1}^{K} P(C_i) = 1
\]

\[
\text{likelihood, } P(x | C_i)
\]

\[
\text{posterior, } P(C_i | x) = \frac{P(x | C_i) P(C_i)}{P(x)}
\]

Bayes’ classification: pick the class with the highest posterior probability:

\[
\text{choose } C_i : P(C_i | x) = \max_k P(C_k | x)
\]
(4) \textbf{Losses & Risks}

But decisions may not be equally good or costly. Need to allow for different gain & loss.

Action $\alpha_i$ is decision to classify input to $C_i$.

$\ell_{ik}$ loss incurred for taking action $\alpha_i$ if class is $C_k$.

Expected risk:

$$R(\alpha_i|x) = \sum_{k=1}^{K} \ell_{ik} P(C_k|x) \tag{4.1}$$

Choose action with minimal risk:

choose $\alpha_i$ if $R(\alpha_i|x) = \min_k R(\alpha_k|x)$.

$\ell$ action, $\alpha_i \in \{1, \ldots, K\}$ action assigned to $C_i$.

Zero-one loss

$\ell_{ik} = \{ 0, \quad \text{if } i=k, \}

R(\alpha_i|x) = \sum_{k=1}^{K} \ell_{ik} P(C_k|x) \tag{4.2}

= \sum_{k \neq i} P(C_k|x) = 1 - P(C_i|x) \tag{4.3}

To minimize risk (zero-one loss) we choose

the most probable class $P(C_i|x) = \max_k P(C_k|x)$.  

(5) In some applications — e.g. medical diagnosis — wrong decisions may have a very high cost — and we may require a more complex decision:

New action — reject/doubt $x_{k+1}$.

$$\tau_{ik} = \begin{cases} 
0, & \text{if } i = k \\
\tau, & \text{if } i = k+1 \\
1, & \text{otherwise}
\end{cases}$$

Risk of reject

$$R(d_{k+1}|x) = \sum_{k=1}^{K} \tau \cdot P(C_k|x) = \tau$$

Risk of class $i$

$$R(d_i|x) = \sum_{k \neq i} \tau \cdot P(C_k|x) = 1 - P(C_i|x).$$

Optimal Decision Rule:

Choose $C_i$ if $R(d_i|x) < R(d_k|x) \forall k \neq i$

reject $d_i$ if $R(d_{k+1}|x) < R(d_i|x)$, $i = 1, K$

Given simple loss function.

choose $C_i$ if $P(C_i|x) > P(C_k|x) \forall k \neq i$

$P(C_i|x) > 1 - \alpha$

reject, otherwise.

$\alpha < \lambda < 1$, $\lambda = 0$, always reject.

$\lambda = 1$, never reject.
Discriminant Functions.

Classification can be seen as implementing a set of discriminant functions $g_i(x), i=1,...,K$. Choose $C_i$, if $g_i(x) = \max_k g_k(x)$.

$$g_i(x) = -R(C_i|x)$$

The maximum discriminant function corresponds to minimum risk.

With zero-one loss function,

$$g_i(x) = P(C_i|x)$$

or

$$g_i(x) = P(x|C_i)P(C_i)$$

ignoring normalization term $P(x)$.

These divide the feature space into $K$ decision regions $R_1,...,R_K$.

$$R_i = \{ x \mid g_i(x) = \max_k g_k(x) \}$$

separated by decision boundaries.

(Where discriminants are tied).

For two classes,

single discriminant $g(x) = g_1(x) - g_2(x)$

choose $\{ C_1 \mid g(x) > 0 \}$

$C_2$ otherwise.
Utility Theory

Utility function $U(x)$.

Expected utility $EU(x) = \sum \frac{1}{S} U(x_k) P(S_k|x)$

States $S$. 

Maximizing utility - or minimize risk.

Value of Information.

Medical diagnosis - many tests can be applied, but some are expensive.

Measuring pulse is cheap.

Blood test is costly.

But blood test may give more information.

If we observe $x$, then expected utility is $EU(x) = \max \sum \frac{1}{S} U(x_k) P(S_k|x)$

If we observe new feature $z$.

$EU(x,z) = \max \sum \frac{1}{S} U(x_k) P(S_k|x,z)$

If $EU(x,z) > EU(x)$, then $z$ is useful.

Difference is the value of information.

(But need some way to estimate the value of $z$).
Bayesian Networks - optional topic

\[ p(x_1 \ldots x_n) = \prod_i p(x_i | x_{\text{parent}(i)}) \]

Naive Bayes classifier (ignores dependencies)

\[ p(x | c_i) = \prod_{j=1}^{k} p(x_j | c_i) \]

Usually there is more structure

Influence Diagrams

Association Rules:

Confidence \((x \rightarrow y) = \frac{P(y|x)}{P(x)} = \frac{P(x,y)}{P(x)}\)

Support of association rule

\[ \text{Support}(x,y) = P(x \cap y) \]