Alpaydin chp 3.

Note Title 3/26/2008

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 $\quad X=1$ heads, $X=0$ tails.

$p(X=1) = p_0$ & $p(X=0) = 1 - P(X=1) = 1 - p_0$.

If we know $p(x)$.

Then we can predict:

i.e if $p_0 > \frac{1}{2}$, predict heads

otherwise, predict tails

If we do not know $p(x)$, then learn it from training samples

$$\hat{p}_0 = \frac{\#\text{ tosses which are heads}}{\#\text{ tosses}}$$

(2)

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 $P(C \mid X_1, X_2)$ and get a new customer with (x_1, x_2)

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

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

Prob of error $= 1 - \max\{P(C=1 \mid x_1, x_2), P(C=0 \mid x_1, x_2)\}$

How to calculate $P(C \mid \underline{x})$? $\underline{x} = (x_1, x_2)$

Bayes Rule $P(C \mid \underline{x}) = \dfrac{p(C)\, p(\underline{x} \mid C)}{p(\underline{x})}$

$p(C)$ ~ prior probability (before observing data)

$p(\underline{x} \mid C)$ ~ class likelihood.

$p(\underline{x})$ ~ evidence

(3)

Classification (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 $C_i : i = 1 \ldots K$

prior, $p(C_i) \geq 0$, $\sum_{i=1}^{K} p(C_i) = 1$

likelihood, $p(\underline{x} | C_i)$,

posterior, $p(C_i | \underline{x}) = \frac{p(\underline{x} | C_i) p(C_i)}{p(\underline{x})}$

Bayes' classifier pick the class with the highest posterior probability:

choose C_i : if $p(C_i | \underline{x}) = \max_k p(C_k | \underline{x})$

(4)

Losses & Risks

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

Action α_i is decision to classify input to C_i.

λ_{ik} loss incurred for taking action α_i if class is C_k.

Expected risk:

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

Choose action with minimal risk:

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

K actions, α_i s.t. $i=1,..K$ α_i assign to C_i

zero-one loss $\lambda_{ik} = \begin{cases} 0, & \text{if } i=k \\ 1, & \text{if } i \neq k. \end{cases}$

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

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

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 α_{K+1}.

$$\lambda_{ik} = \begin{cases} 0, & \text{if } i = k \\ \lambda, & \text{if } i = K+1 \\ 1, & \text{otherwise} \end{cases} \qquad 0 < \lambda < 1$$

Risk of reject

$$R(\alpha_{K+1}|\underline{x}) = \sum_{k=1}^{K} \lambda P(C_k|\underline{x}) = \lambda$$

Risk of class i

$$R(\alpha_i|\underline{x}) = \sum_{k \neq i}^{K} P(C_k|\underline{x}) = 1 - P(C_i|\underline{x}).$$

Optimal Decision Rule:

Choose C_i if $R(\alpha_i|\underline{x}) < R(\alpha_k|\underline{x}) \ \forall k \neq i$ & $R(\alpha_i|\underline{x}) < R(\alpha_{K+1}|\underline{x})$.

reject if $R(\alpha_{K+1}|\underline{x}) < R(\alpha_i|\underline{x}), \ i = 1, K$

Given simple loss function.

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

$P(C_i|\underline{x}) > 1 - \lambda$

reject, otherwise.

$0 < \lambda < 1$, if $\lambda = 0$, always reject

$\lambda = 1$, never reject.

Discriminant Functions.

Classification can be seen as implementing a set of discriminant functions $g_i(\underline{x})$, $i = 1, \dots K$ s.t.

choose C_i, if $g_i(\underline{x}) = \max_k g_k(\underline{x})$

$$g_i(x) = -R(\alpha_i | x)$$

maximum discriminant function corresponds to minimum risk.

with zero-one loss function,

$$g_i(\underline{x}) = P(C_i | \underline{x})$$

or $$g_i(\underline{x}) = p(\underline{x} | C_i) P(C_i)$$

ignoring normalization constant term $p(x)$.

These divide the feature space into K decision regions $\mathcal{R}_1 \dots \mathcal{R}_K$

$$\mathcal{R}_i = \{ \underline{x} \mid g_i(\underline{x}) = \max_k g_k(\underline{x}) \}$$

separated by decision boundaries.

(where discriminants are tied).

For two classes,

single discriminant $g(\underline{x}) = g_1(\underline{x}) - g_2(\underline{x})$

choose $\begin{cases} C_1 & \text{if } g(\underline{x}) > 0 \\ C_2 & \text{otherwise} \end{cases}$

17)

Utility Theory

optimistic version of decision theory.

Utility function U_{ik}.

Expected utility $EU(\alpha_i|\underline{x}) = \sum_k U_{ik} P(S_k|\underline{x})$

States S_k

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 $\underline{x}$, then expected utility is

$$EU(\underline{x}) = \max_i \sum_k U_{ik} P(S_k|\underline{x})$$

If we observe new feature z

$$EU(\underline{x}, z) = \max_i \sum_k U_{ik} P(S_k|\underline{x}, z)$$

If $EU(\underline{x}, z) > EU(\underline{x})$, then z is useful.

Difference is the value of information

(But need some way to estimate the value of z).

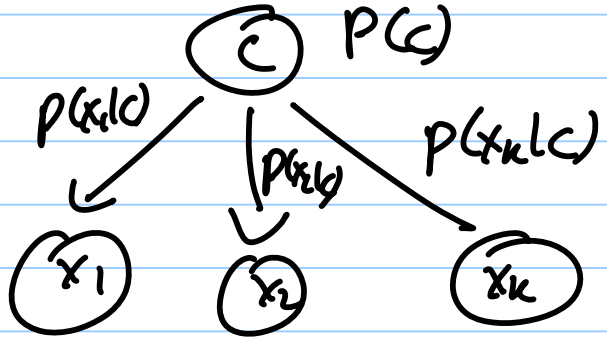
(8)

Bayesian Networks – optional topic

$$p(x_1 \dots x_n) = \prod_i p(x_i \mid X_{p(i)})$$

* parents

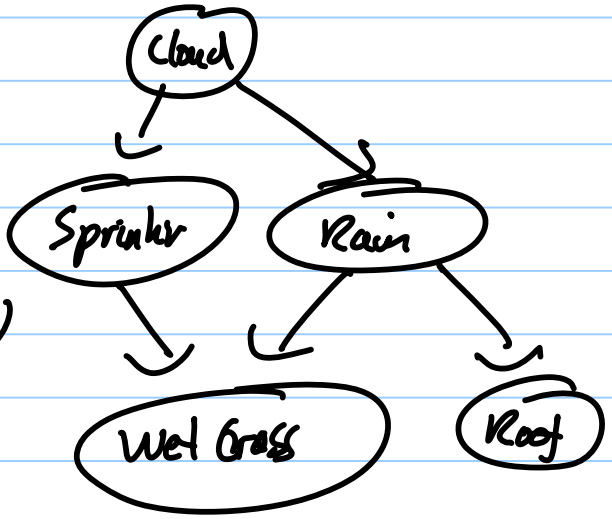
Naive Bayes classifier (ignores dependencies)



$$p(\underline{x} \mid C_i) = \prod_{j=1}^{k} p(x_j \mid C_i)$$

Usually there is more structure

Influence Diagrams



Association Rules:

Confidence $(X \to Y) = P(Y \mid X) = \frac{P(X,Y)}{P(X)}$

Support of association rule

Support $(X, Y) = P(X, Y)$