Computational Intelligence: Methods and Applications

Lecture 12

Bayesian decisions: foundation of learning

Włodzisław Duch

SCE, NTU, Singapore

Google: Duch

Learning

- Learning from data requires a model of data.
- Traditionally parametric models of different phenomena were developed in science and engineering; parametric models are easy to interpret, so if the phenomenon is simple enough and a theory exist construct a model and use algorithmic approach.
- Empirical non-parametric modeling is data driven, goal oriented. It dominates in biological systems. Learn from data!
- Given some examples = training data, create a model of data that answers specific question, estimating those characteristics of the data that may be useful to make future predictions.
- Learning = estimate parameters of the (non-parametric) model; paradoxically, non-parametric models have a lot of parameters.
- Many other approaches to learning exist, but no time to explore ...

Probability

To talk about prediction errors probability concepts are needed.

Samples $X \in \mathcal{X}$ divided into K categories, called classes $\omega_1 \dots \omega_K$

More general, ω_i is a state of nature that we would like to predict.

$P_k = P(\omega_k)$, *a priori* (unconditional) probability of observing $X \in \omega_k$

$$\sum_{k=1}^{K} P_k = 1; \quad P_k = \frac{N(\omega_k)}{N}$$

If nothing else is known than one should predict that a new sample X belongs to the majority class:

$$\mathbf{X} \in \omega_c; \quad c = \arg\max_k P_k$$

Majority classifier: assigns all new X to the most frequent class.

Example: weather prediction system – the weather tomorrow will be the same as yesterday (high accuracy prediction!).

Conditional probability

Predictions should never be worse than for the majority classifier! Usually class-conditional probability is also known or may easily be measured, the condition here is that $X \in \omega_k$

$$P_k(\mathbf{X}) = P(\mathbf{X}|\omega_k) = P(\mathbf{X}|C = \omega_k)$$

Joint probability of observing X from ω_k

$$P(\mathbf{X}, \omega_k) = P(\mathbf{X}|\omega_k) P(\omega_k)$$

Is the knowledge of conditional probability sufficient to make predictions?

No! We are interested in the *posterior P*.

$$P(\omega_k|\mathbf{X}) = P(\mathbf{X}, \omega_k) / P(\mathbf{X})$$

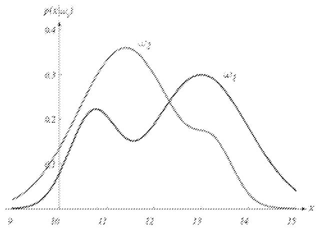

Fig. 2.1, from Duda, Hart, Stork, Pattern Classification (Wiley).

Bayes rule

Posterior conditional probabilities are normalized:

$$\sum_{k=1}^{K} P(\omega_k \mid \mathbf{X}) = 1$$

Bayes rule for 2 classes is derived from this:

$$P(\omega_i, \mathbf{X}) = P(\omega_i \mid \mathbf{X}) P(\mathbf{X})$$
$$= P(\mathbf{X} \mid \omega_i) P(\omega_i)$$

$P(\mathbf{X})$ is an unconditional probability of selecting sample **X**; usually it is just $1/n$, where n=number of all samples.

For P_1=2/3 and P_2=1/3 previous figure is:

$$P(\omega_i \mid \mathbf{X}) = P(\mathbf{X} \mid \omega_i) P(\omega_i) / P(\mathbf{X})$$

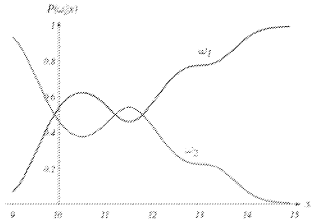

Fig. 2.2, from Duda, Hart, Stork, Pattern Classification (Wiley).

Bayes decisions

Bayes decision: given a sample X select class 1 if:

$$P(\omega_1 \mid \mathbf{X}) > P(\omega_2 \mid \mathbf{X})$$

Using Bayes rule and multiplying both sides by P(X):

$$P(\mathbf{X} \mid \omega_1) P(\omega_1) > P(\mathbf{X} \mid \omega_2) P(\omega_2)$$

Probability of an error is:

$$P(\varepsilon \mid \mathbf{X}) = \min\left(P(\omega_1 \mid \mathbf{X}), P(\omega_2 \mid \mathbf{X})\right)$$

Average error is:

$$P(\varepsilon) = E\left[P(\varepsilon \mid \mathbf{X})\right] = \int_{-\infty}^{+\infty} P(\varepsilon \mid \mathbf{X}) P(\mathbf{X}) d\mathbf{X}$$

Bayes decision rule minimizes average error selecting smallest $P(\varepsilon \mid \mathrm{X})$

Likelihood

On a finite data sample given for training the error is:

$$P(\varepsilon) = \sum_{\mathbf{X}} P(\varepsilon \mid \mathbf{X})$$

The assumption here is that the P(X) is reflected in the frequency of samples for different X.

Bayesian approach to learning: use data to model probabilities.
Bayes decision depends on the **likelihood ratio**:

$$P(\mathbf{X} \mid \omega_1) P(\omega_1) > P(\mathbf{X} \mid \omega_2) P(\omega_2)$$

$$\Lambda(\mathbf{X}) = \frac{P(\mathbf{X} \mid \omega_1)}{P(\mathbf{X} \mid \omega_2)} > \frac{P(\omega_2)}{P(\omega_1)}$$

For equal *a priori* probabilities class conditional probabilities decide.

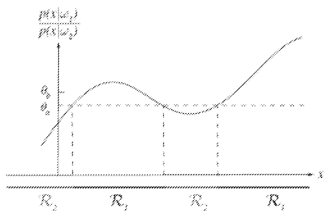

Fig. 2.3, from Duda, Hart, Stork, Pattern Classification (Wiley).

2D decision regions

For Gaussian distribution of class conditional probabilities:

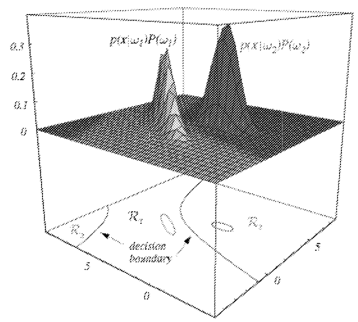

Fig. 2.6, from Duda, Hart, Stork, Pattern Classification (Wiley).

Decision boundaries in 2D are hyperbolic, decision region R_2 is disconnected. The ellipsis show high constant values of P_k(X).

Example

Let ω_1 be the state of nature called "dengue",
and ω_2 the opposite, no dengue.
Let prior probability for people in Singapore be $P(\omega_1)=0.1\%$
Let test T be accurate in 99%, so that the positive outcome of the test for people with dengue is $P(T=+|\omega_1) = 0.99$, and negative for healthy people is also $P(T=-|\omega_2) = 0.99$.

What is the chance that you have dengue if your test is positive?

What is the probability $P(\omega_1|T=+)$?
$P(T=+) = P(\omega_1,T=+)+P(\omega_2,T=+) = 0.99*0.001+0.01*0.999=0.011$
$P(\omega_1|T=+)=P(T=+|\omega_1)P(\omega_1)/P(T=+) = 0.99*0.001/0.011 = 0.09$, or 9%

Use this calculator to check:
http://members.aol.com/johnp71/bayes.html

Statistical theory of decisions

Decisions carry risk, costs and losses.
Consider general decision procedure:

$\{\omega_1 .. \omega_K\}$, states of nature
$\{\alpha_1 .. \alpha_a\}$, actions that may be taken
$\lambda(\omega_i, \alpha_j)$, cost, loss or risk associated with action α_j in state ω_i

Example: classification decisions

$$\hat{C}: \mathbf{X} \rightarrow \{\omega_1 .. \omega_K, \omega_D, \omega_O\},$$

Action α_i is assigning to vector X a class label 1 .. *K*, or
ω_D – no confidence in classification, reject/leave sample as unclassified
ω_O – outlier, untypical case, perhaps a new class (used rarely).

Errors and losses

Unconditional probability of wrong (non-optimal) action (decision), if the true state is ω_k, and prediction was wrong:

$$P_\varepsilon(\omega_k) = P\left\{\hat{C}(\mathbf{X}) \neq \omega_k \wedge \hat{C}(\mathbf{X}) \in \{\omega_1..\omega_K\} \mid C = \omega_k\right\}$$

No action (no decision), or rejection of sample X if the true class is ω_k, has probability:

$$P_D(\omega_k) = P\left\{\hat{C}(\mathbf{X}) = \omega_D \mid C = \omega_k\right\}$$

Assume simplest 0/1 loss function: no cost if optimal decision is taken, identical costs for all errors and some costs for no action (rejection):

$$\lambda_{kl} = \lambda(\omega_k, \omega_l) = \begin{cases} 0 \text{ if} & k = l \\ 1 \text{ if} & k \neq l, l \in \{1 \ldots K\} \\ \varepsilon_d \text{ if} & l = D \end{cases}$$

... and risks

Risk of the decision making procedure $\hat{C}$ for class ω_k, with $\omega_{K+1}=\omega_D$

$$R(\hat{C}, \omega_k) = E\left[\lambda(\omega_k, \hat{C}(\mathbf{X})) \mid C = \omega_k\right] = \sum_{l=1}^{K+1} \lambda(\omega_k, \omega_l) P_{kl} = P_\varepsilon(\omega_k) + \varepsilon_d P_D(\omega_k)$$

where P_{kl} are elements of the confusion matrix **P**:

$$P\left\{\hat{C}(\mathbf{X}) = \omega_l \mid C = \omega_k\right\} = \mathbf{P} = \begin{bmatrix} P_{11} & P_{12} & \ldots & P_{1K+1} \\ P_{21} & P_{22} & \ldots & P_{2K+1} \\ \ldots & \ldots & \ldots & \ldots \\ P_{K1} & P_{K2} & \ldots & P_{KK+1} \end{bmatrix}$$

Note that rows of **P** correspond to the true ω_k classes, and columns to the predicted ω_l classes, $l = 1..K+1$ or classifier's decisions.

... and more risks

Trace of the confusion matrix:

$$A = \mathrm{Tr}\,\mathbf{P} = \sum_{i=1}^{K} P_{ii}$$

is equal to accuracy of the classifier, ignoring costs of mistakes.

Total risk of the decision procedure $\hat{C}$:

$$R\left(\hat{C}\right) = \sum_{k=1}^{K} P(\omega_k) R\left(\hat{C}, \omega_k\right) = \sum_{k=1}^{K} \sum_{l=1}^{K+1} \lambda_{kl} P_{kl} P(\omega_k)$$

$$= \sum_{k=1}^{K} P(\omega_k)\left(P_\varepsilon(\omega_k) + \varepsilon_d P_D(\omega_k)\right)$$

For special case of costs of mistakes =1 and rejection $=\varepsilon_d$

Conditional risk of assigning sample X to class ω_k is:

$$R(\omega_k \mid \mathbf{X}) = \sum_{j=1}^{K+1} \lambda_{kj} P\left(\omega_j \mid \mathbf{X}\right)$$