Bayesian Decision Theory

Learning as Bayesian Inference

- Formulate the learning task as a process of probabilistic inference
  - Inference step: determine P(x,t) from data
  - Decision step: for given x, determine optimal t.
- *Bayesian Decision Theory*
  - A fundamental statistical approach to the problem of pattern recognition and machine learning
- Bayesian framework provides a sound probabilistic basis for understanding many learning algorithms and designing new algorithms

1

A Classification Problem

- The sea bass/salmon example:
  a fish-packing plant wants to automate the process of sorting incoming fish on a conveyor belt according to species
- State of nature, ω (the category of the fish)
  - $\omega = \omega_1$ *sea bass,* $\omega = \omega_2$ *salmon*
  - State of nature is a random variable
- We assume that there is some *a priori probability* $P(\omega)$, *Prior*
  - The catch of salmon and sea bass is equally probable?
  - Prior probabilities reflect our prior knowledge

2

- Measure a feature value *X, say* length : continuous random variable
- Different fish will yield different length readings:
  - *class–conditional probability* density function $p(x \mid \omega)$
- $p(x \mid \omega_1)$ and $p(x \mid \omega_2)$ describe the difference in length between populations of sea bass and salmon

3

Bayes Rule

- Suppose we measure the length of a fish and discover that its value is *x*
- *A posteriori probability (or posterior) $P(\omega_j \mid x)$:* the probability of the state of nature given that feature value *x* has been measured
- $P(\omega_j \mid x) = p(x \mid \omega_j) \cdot P(\omega_j) / p(x)$
- *Posterior = (Likelihood x Prior) / Evidence*

4

- Decision given the posterior probabilities

  x is an observation for which:

  if $P(\omega_1 \mid x) > P(\omega_2 \mid x)$ ⇨ True state of nature $= \omega_1$
  if $P(\omega_1 \mid x) < P(\omega_2 \mid x)$ ⇨ True state of nature $= \omega_2$

5

Justification

- Whenever we observe a particular x, the probability of error is :

  $P(error \mid x) = P(\omega_1 \mid x)$ *if we decide* ω_2

  $P(error \mid x) = P(\omega_2 \mid x)$ *if we decide* ω_1

- Minimizing the probability of error
  - Decide ω_1 *if* $P(\omega_1 \mid x) > P(\omega_2 \mid x)$; otherwise decide ω_2
- **Bayes decision rule**
- Bayesian classifier is optimal in that it is guaranteed to minimize the probability of misclassification

6

- Equivalent decision rule

  Decide ω_1 *if* $p(x|\omega_1)\, P(\omega_1) > p(x|\omega_2)\, P(\omega_2)$; otherwise decide ω_2
- If $P(\omega_1) = P(\omega_2)$, the decision is based entirely on the likelihoods $p(x|\omega_1)$ and $p(x|\omega_2)$
- Bayes classification rule combines the effect of the two terms optimally - so as to yield minimum error classification.

7

Bayesian Decision Theory

- Generalization of the preceding ideas
  - Use of more than one feature
  - Use more than two states of nature
  - Allowing actions and not only decide on the state of nature
  - Introduce a loss of function which is more general than the probability of error

8

- Feature vector $\boldsymbol{x}$ in R^d feature space (attributes)
- Let $\{\omega_1, \omega_2, \ldots, \omega_c\}$ be the set of c states of nature (or "categories")
- Let $\{\alpha_1, \alpha_2, \ldots, \alpha_a\}$ be the set of possible actions
  - Typically α_i : decide ω_i
  - Allowing actions other than classification primarily allows the possibility of rejection
  - Refusing to make a decision in close or bad cases!

9

- The loss function states how costly each action taken is
- Situations in which some kinds of classification mistakes are more costly than others
- The simplest case: all errors are equally costly
- Let $\lambda(\alpha_i \mid \omega_j)$ be the loss incurred for taking action α_i when the state of nature is ω_j

10

- Suppose we observe ***x***
- The expected loss associated with taking action α_i called *conditional risk*

$$R(\alpha_i \mid x) = \sum_{j=1}^{j=c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid x)$$

- We minimize loss or overall risk by selecting the action that minimizes the conditional risk

Select the action α_i for which $R(\alpha_i \mid x)$ is minimum

- Bayes decision rule
- Best performance that can be achieved!

11

- Two-category classification

  α_1 : deciding ω_1

  α_2 : deciding ω_2

  $\lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$

loss incurred for deciding ω_i when the true state of nature is ω_j

Conditional risk:

$$R(\alpha_1 \mid x) = \lambda_{11}P(\omega_1 \mid x) + \lambda_{12}P(\omega_2 \mid x)$$
$$R(\alpha_2 \mid x) = \lambda_{21}P(\omega_1 \mid x) + \lambda_{22}P(\omega_2 \mid x)$$

12

Our rule is the following:

if $R(\alpha_1 \mid x) < R(\alpha_2 \mid x)$

action α_1: "decide ω_1" is taken

This results in the equivalent rule :
decide ω_1 if:

$$(\lambda_{21} - \lambda_{11})\, p(x \mid \omega_1)\, P(\omega_1) >$$
$$(\lambda_{12} - \lambda_{22})\, p(x \mid \omega_2)\, P(\omega_2)$$

and decide ω_2 otherwise

13

- Reasonable assumption: the loss incurred for making an error is greater than that for being correct, λ_{21}- λ_{11} and λ_{12}- λ_{22} are positive
- Likelihood ratio:

The preceding rule is equivalent to the following rule:

$$if \ \frac{p(x \mid \omega_1)}{p(x \mid \omega_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)}$$

Then take action α_1 (decide ω_1)
Otherwise take action α_2 (decide ω_2)

14

Minimum-Error-Rate Classification

- We only care about making correct classification
- Actions are decisions on classes, α_i : decide ω_i
  If action α_i is taken and the true state of nature is ω_j then: the decision is correct if $i = j$ and in error if $i \neq j$
- Seek a decision rule that minimizes the *probability of error* which is the *error rate*

15

- Introduction of the zero-one loss function:

$$\lambda(\alpha_i \mid \omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \qquad i, j = 1, \ldots, c$$

Therefore, the conditional risk is:

$$R(\alpha_i \mid x) = \sum_{j=1}^{j=c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid x)$$

$$= \sum_{j \neq i} P(\omega_j \mid x) = 1 - P(\omega_i \mid x)$$

"The risk corresponding to this loss function is the average probability error"

16

- Minimize the risk requires maximize $P(\omega_i \mid x)$ (since $R(\alpha_i \mid x) = 1 - P(\omega_i \mid x)$)

- Bayes decision rule for minimum error rate
  - Decide ω_i if $P(\omega_i \mid x) > P(\omega_j \mid x) \;\; \forall j \neq i$

17

Summary of Bayesian recipe for classification

$$R(\alpha_i \mid x) = \sum_{j=1}^{j=c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid x)$$

- *Select the action α_i for which $R(\alpha_i \mid x)$ is minimum*

- For minimum error rate
  - Decide ω_i *if* $P(\omega_i \mid x) > P(\omega_j \mid x)$ $\forall j \neq i$

18

Example: The Normal Density

- Univariate normal (Gaussian) density

$$p(x) = N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right],$$

Where:

μ = mean (or expected value) of x

σ^2 = variance, σ the standard deviation

- The expected value (mean, average) of x

$$\mu = E[x] = \int_{-\infty}^{\infty} x p(x) dx$$

- The variance

$$\mathrm{Var}[x] = \sigma^2 = E[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx$$

19

Example

Select the optimal decision where:

$\{\omega_1, \omega_2\}$

$p(x \mid \omega_1)$ ⟹ N(1, 1) (Normal distribution)

$p(x \mid \omega_2)$ ⟹ N(2, 1)

$P(\omega_1) = 2/3, \; P(\omega_2) = 1/3$

- For Minimum error rate
  - Decide ω_1 *if* $x < 1.5 + ln2$
- If $\lambda = \begin{bmatrix} 0 & 1 \\ 2 & 0 \end{bmatrix}$
  - Decide ω_1 *if* $x < 1.5 + ln4$

20

Discriminant Functions

- One way to represent pattern classifiers is in terms of discriminant functions
- The multi-category case
  - Set of discriminant functions $g_i(x), \; i = 1, \ldots, c$
  - The classifier assigns a feature vector x to class ω_i if:

$$g_i(x) > g_j(x) \;\; \forall j \neq i$$

21

Bayes Classifier

- *Let* $g_i(x) = - R(\alpha_i \mid x)$
  (max. discriminant corresponds to min. risk!)
- For the minimum error rate, we take
  $$g_i(x) = P(\omega_i \mid x)$$
  (max. discrimination corresponds to max. posterior!)

  $$g_i(x) \equiv p(x \mid \omega_i)\, P(\omega_i)$$

  $$g_i(x) = \ln p(x \mid \omega_i) + \ln P(\omega_i)$$
- Equivalent discriminants, some can be simpler to compute than others

22

- The two-category case
  - Instead of using two discriminant functions g_1 and g_2
    Let $g(x) \equiv g_1(x) - g_2(x)$
    Decide ω_1 *if* $g(x) > 0$; Otherwise decide ω_2
  - Minimum-error-rate discriminant

$$g(x) = P(\omega_1 \mid x) - P(\omega_2 \mid x)$$

$$g(x) = \ln \frac{p(x \mid \omega_1)}{p(x \mid \omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

23

Example

$p(x \mid \omega_1)$ ⟹ N(1, 1) (Normal distribution)
$p(x \mid \omega_2)$ ⟹ N(2, 1)
$P(\omega_1) = 2/3,\ P(\omega_2) = 1/3$

- For Minimum error rate
  - Decide $\boldsymbol{\omega_1}$ *if* $x < 1.5 + ln2$
  - $g(x) = 1.5 + ln2 - x$

Example: The Normal Density

- Multivariate normal density $p(x) \sim N(\mu, \Sigma)$

$$p(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^t \Sigma^{-1}(x-\mu)\right]$$

where:
$x = (x_1, x_2, \ldots, x_d)^t$ (t stands for the transpose vector form)
$\mu = (\mu_1, \mu_2, \ldots, \mu_d)^t$ mean vector
$\Sigma = d*d$ covariance matrix
$|\Sigma|$ and Σ^{-1} are determinant and inverse respectively

- The covariance of x_i *and* x_j

$$\sigma_{ij} = E[(x_i - \mu_i)(x_j - \mu_j)] = \int_{-\infty}^{\infty}(x_i - \mu_i)(x_j - \mu_j)p(x)dx$$

- If x_i *and* x_j are independent, then $\sigma_{ij} = 0$

Discriminant Functions for the Normal Density

- Generative model: multivariate normal $p(x \mid \omega_i) \sim N(\mu_i, \Sigma_i)$
- The minimum error-rate classification can be achieved by the discriminant function

$g_i(x) = \ln p(x \mid \omega_i) + \ln P(\omega_i)$

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

26

- Case Σ_i = arbitrary
  - The covariance matrices are different for each category

$$g_i(x) = x^t W_i x + w_i^t x + w_{i0}$$

where :

$$W_i = -\frac{1}{2}\Sigma_i^{-1}$$

$$w_i = \Sigma_i^{-1}\mu_i$$

$$w_{i0} = -\frac{1}{2}\mu_i^t \Sigma_i^{-1}\mu_i - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

(The decision surfaces are hyperquadrics which are: hyperplanes, pairs of hyperplanes, hyperspheres, hyperellipsoids, hyperparaboloids, hyperhyperboloids)

27

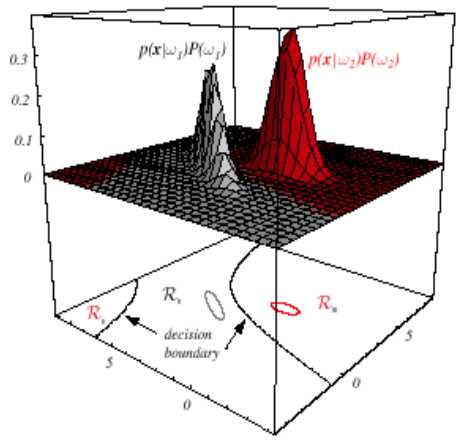

FIGURE 2.6. In this two-dimensional two-category classifier, the probability densities are Gaussian, the decision boundary consists of two hyperbolas, and thus the decision region $\mathcal{R}_2$ is not simply connected. The ellipses mark where the density is $1/e$ times that at the peak of the distribution. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

28

- Case $\Sigma_i = \Sigma$ (covariance of all classes are identical but arbitrary)
  - Ignore terms independent of *i*

$$g_i(x) = w_i^t x + w_{i0}$$

where:

$$w_i = \Sigma^{-1}\mu_i; \quad w_{i0} = -\frac{1}{2}\mu_i^t \Sigma^{-1}\mu_i + \ln P(\omega_i)$$

Linear discriminant functions!

29

Summary of Bayesian recipe for classification

• The Bayesian recipe is simple, optimal, and in principle, straightforward to apply

- We could design an optimal classifier if we knew:
  - $P(\omega_i)$ (priors)
  - $p(x \mid \omega_i)$ (class-conditional densities)
- Unfortunately, we rarely have this complete information!
- We have some knowledge and training data
  - training data set $\{(\mathbf{x}_i, \omega_i)\}$
- Use the samples to estimate the unknown probability distributions

30

Inference and Decision

- *Generative approach:*
  Model $P(\omega_i, x) = p(x \mid \omega_i) P(\omega_i)$
  Use Bayes' theorem to obtain $P(\omega_i \mid x)$
- *Discriminative approach:*
  Model $P(\omega_i \mid x)$ directly
- Directly model discrimination functions

31