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
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, $\omega$ (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

Suppose we have to make a decision about the type of fish that will appear next without seeing it.

- A seemingly logical decision rule with only the prior information  
  - Decide $\omega_1$ if $P(\omega_1) > P(\omega_2)$; otherwise decide $\omega_2$

- The probability of error is the smaller of $P(\omega_1)$ and $P(\omega_2)$
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.

Bayes Rule

Suppose we measure the length of a fish and discover that its value is $x$.

A posteriori probability (or posterior) $P(\omega \mid x)$: the probability of the state of nature given that feature value $x$ has been measured.

$P(\omega_j \mid x) = p(x \mid \omega_j) \cdot P(\omega_j) / p(x)$
• Decision given the posterior probabilities

x is an observation for which:

if $P(\omega_1 | x) > P(\omega_2 | x) \implies$ True state of nature $= \omega_1$

if $P(\omega_1 | x) < P(\omega_2 | x) \implies$ True state of nature $= \omega_2$

Justification

• Whenever we observe a particular x, the probability of error is:
  
  $P(\text{error} | x) = P(\omega_1 | x)$ if we decide $\omega_2$
  $P(\text{error} | x) = P(\omega_2 | x)$ if we decide $\omega_1$

• Minimizing the probability of error
  
  Decide $\omega_i$ if $P(\omega_i | x) > P(\omega_2 | x)$; otherwise decide $\omega_2$

• A decision rule will divide feature space into decision regions
  
  $R_i$ means assign x to $\omega_i$

• The regions are separated by decision boundaries
Optimality of Bayes Decision Rule

\[ P(\text{error}) = P(x \in R_1, \omega_1) + P(x \in R_2, \omega_1) \]
\[ = \int_{R_1} p(\omega_1 | x) p(x) dx + \int_{R_2} p(\omega_1 | x) p(x) dx \]

- Minimizing the probability of error
  - Decide \( \omega_1 \) if \( P(\omega_1 | x) > P(\omega_2 | x) \); otherwise decide \( \omega_2 \)

- Bayes decision rule

- Bayesian classifier is optimal in that it is guaranteed to minimize the probability of misclassification

- Equivalent decision rule
  Decide \( \omega_1 \) if \( p(x | \omega_1) P(\omega_1) > p(x | \omega_2) P(\omega_2) \); otherwise decide \( \omega_2 \)

- 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.
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 function which is more general than the probability of error

• Feature vector $\mathbf{x}$ in $\mathbb{R}^d$ 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 $\alpha_i$: decide $\omega_i$
    • Allowing actions other than classification primarily allows the possibility of rejection
    • Refusing to make a decision in close or bad cases!
• 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 $\alpha_i$ when the state of nature is $\omega_j$

• Suppose we observe $x$

• The expected loss associated with taking action $\alpha_i$ called conditional risk

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

• We minimize overall risk (expected loss associated with a decision rule) by selecting the action that minimizes the conditional risk

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

• Bayes decision rule

• Best performance that can be achieved!
• Two-category classification

\[ \alpha_1 : \text{deciding } \omega_1 \]
\[ \alpha_2 : \text{deciding } \omega_2 \]
\[ \lambda_{ij} = \lambda(\alpha_i | \omega_j) \]

loss incurred for deciding \( \omega_i \) when the true state of nature is \( \omega_j \)

Conditional risk:

\[ R(\alpha_1 | x) = \lambda_{11} P(\omega_1 | x) + \lambda_{12} P(\omega_2 | x) \]
\[ R(\alpha_2 | x) = \lambda_{21} P(\omega_1 | x) + \lambda_{22} P(\omega_2 | x) \]

Our rule is the following:

if \( R(\alpha_1 | x) < R(\alpha_2 | x) \)

action \( \alpha_1 : \) “decide \( \omega_1 \)” is taken

This results in the equivalent rule:

decide \( \omega_1 \) if:

\[ (\lambda_{21} - \lambda_{11}) p(x | \omega_i) P(\omega_i) > (\lambda_{12} - \lambda_{22}) p(x | \omega_2) P(\omega_2) \]

and decide \( \omega_2 \) otherwise
• Reasonable assumption: the loss incurred for making an error is greater than that for being correct, $\lambda_{21} - \lambda_{11}$ and $\lambda_{12} - \lambda_{22}$ are positive

• Likelihood ratio:

The preceding rule is equivalent to the following rule:

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

Then take action $\alpha_1$ (decide $\omega_1$)
Otherwise take action $\alpha_2$ (decide $\omega_2$)

---

Minimum-Error-Rate Classification

• A special case: we only care about making correct classification

• Actions are decisions on classes, $\alpha_i : \text{decide } \omega_i$
  If action $\alpha_i$ is taken and the true state of nature is $\omega_j$ 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
• Introduction of the zero-one loss function:

\[
\lambda(\omega_i | \omega_j) = \begin{cases} 
0 & i = j \\
1 & i \neq j 
\end{cases} 
i, j = 1, \ldots, c
\]

Therefore, the conditional risk is:

\[
R(\omega_i | x) = \sum_{j=1}^{c} \lambda(\omega_i | \omega_j) P(\omega_j | x)
\]

\[
= \sum_{j=1}^{c} P(\omega_j | x) = 1 - P(\omega_i | x)
\]

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

• Minimize the risk requires maximize \( P(\omega_i | x) \)

(since \( R(\omega_i | x) = 1 - P(\omega_i | x) \))

• Bayes decision rule for minimum error rate

• Decide \( \omega_i \) if \( P(\omega_i | x) > P(\omega_j | x) \) \( \forall j \neq i \)
Summary of Bayesian recipe for classification

\[ R(\alpha_i | x) = \sum_{j=1}^{j=c} \lambda(\alpha_i | \omega_j) P(\omega_j | x) \]

• Select the action \( \alpha_i \) for which \( R(\alpha_i | x) \) is minimum

• For minimum error rate
  * Decide \( \omega_i \) if \( P(\omega_i | x) > P(\omega_j | x) \) \( \forall j \neq i \)

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:
\( \mu = \) mean (or expected value) of \( x \)
\( \sigma^2 = \) variance, \( \sigma \) the standard deviation

• The expected value (mean, average) of \( x \)

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

• The variance

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

Select the optimal decision where:

\{\omega_1, \omega_2\}

\[ p(x | \omega_1) \xrightarrow{\text{N}(1, 1)} \text{Normal distribution} \]

\[ p(x | \omega_2) \xrightarrow{\text{N}(2, 1)} \]

\[ P(\omega_1) = 2/3, \ P(\omega_2) = 1/3 \]

- For Minimum error rate
  - Decide \(\omega_1\) if \(x < 1.5 + \ln 2\)

- If

\[
\lambda = \begin{bmatrix} 0 & 1 \\ 2 & 0 \end{bmatrix}
\]

- Decide \(\omega_1\) if \(x < 1.5 + \ln 4\)

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 \(\omega_i\) if:

\[ g_i(x) > g_j(x) \ \forall j \neq i \]
Bayes Classifier

• Let $g_i(x) = - R(\alpha_i | x)$
  (max. discriminant corresponds to min. risk!)

• For the minimum error rate, we take
  
  $g_i(x) = P(\omega_i | x)$

  (max. discrimination corresponds to max. posterior!)

  $$g_i(x) \equiv p(x | \omega_i) P(\omega_i)$$

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

• Equivalent discriminants, some can be simpler to compute than others

• 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 $\omega_1$ if $g(x) > 0$; Otherwise decide $\omega_2$

  • Minimum-error-rate discriminant

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

  $$g(x) = \ln \frac{p(x | \omega_1)}{p(x | \omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$
Example: The Normal Density

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

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

where:
- $x = (x_1, x_2, \ldots, x_d)'$ (t stands for the transpose vector form)
- $\mu = (\mu_1, \mu_2, \ldots, \mu_d)'$ mean vector
- $\Sigma = d \times d$ covariance matrix
- $|\Sigma|$ and $\Sigma^{-1}$ are determinant and inverse respectively

- The covariance of $x_i$ and $x_j$

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

- 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

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

\[
g_i(x) = -\frac{1}{2} (x - \mu_i)'\Sigma_i^{-1}(x - \mu_i) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_i| + \ln P(\omega_i)
\]
Consider some special cases

- **Case** $\Sigma_i = \sigma^2 I$ (I stands for the identity matrix)
  - The features are independent and each feature has the same variance

ignore terms independent of $i$

$$g_i(x) = w_i^t x + w_i^0$$ (linear discriminant functions!)

where:

$$w_i = \frac{\mu_i}{\sigma^2}; \quad w_i^0 = -\frac{1}{2\sigma^2} \mu_i^t \mu_i + \ln P(\omega_i)$$

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

$$g_i(x) = w_i^t x + w_i^0$$

where:

$$w_i = \Sigma^{-1} \mu_i; \quad w_i^0 = -\frac{1}{2} \mu_i^t \Sigma^{-1} \mu_i + \ln P(\omega_i)$$

- Linear discriminant functions!
• Case $\Sigma_i = \text{arbitrary}$

  • 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)

---

**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)
  • $p(x | \omega_i)$ (the generative model for data)

• Unfortunately, we rarely have this complete information!

• We have some knowledge and training data
  • A training data set $\{(x, \omega_i)\}$

• Use the samples to estimate the unknown probability distributions
Inference and Decision

• **Generative approach:**
  Model $P(\omega, x) = p(x | \omega) P(\omega)$
  Use Bayes’ theorem to obtain $P(\omega | x)$

• ** Discriminative approach:**
  Model $P(\omega_i | x)$ directly

• Directly model discrimination functions