M2. Bayesian Decision Theory (incl. Bayes classifiers)

Manikandan Narayanan

Feb 4-5,10-11,15 2021

PRML Jan-May 2021 (MKN section)

Acknowledgment of Sources

- Slides based on content from related
  - Courses:
    - IITM – Profs. Arun/Harish/Chandra's PRML offerings (slides, quizzes, notes, etc.), Prof. Ravi's "Intro to ML" slides – cited respectively as [AR], [HR], [CC], [BR] in the bottom right of a slide.
    - India – NPTEL PR course by IISc Prof. PS. Sastry (slides, etc.) – cited as [PSS] in the bottom right of a slide.
  - Books:
    - PRML by Bishop. (content, figures, slides, etc.) – cited as [CMB]
    - Pattern Classification by Duda, Hart and Stork. (content, figures, etc.) – [DHS]
    - Mathematics for ML by Deisenroth, Faisal and Ong. (content, figures, etc.) – [DFO]

Outline of Module M2

- M2. Bayesian Decision Theory (incl. Bayes classifiers)
  - M2.0 Introduction/Background (on Probability Theory)
  - **M2.1 Bayesian Decision Theory**
    - **M2.1.0 Decision Theory for Classification/Regression (common defns./notations)**
    - M2.1.1 Decision Theory for Classification (Bayes classifiers)
    - M2.1.2 Decision Theory for Regression (Squared loss, etc.)

M2.1.0 Decision Theory (for classification/regression)

x is feature vector (input), t is target/response (output).

- Inference step
- Determine either $p(t|\mathrm{x})$ or $p(\mathrm{x}, t)$.

- Decision step
- For any given x, determine optimal t.
- Optimality wrt *risk* or *expected loss*; General loss functions are:
  - Classification (t discrete): ***misclassification rate***, loss-matrix based function, etc.
  - Regression (t continuous): ***squared loss***, Minkowski loss, etc.

[CMB]

Notations

- Feature vector $\mathrm{x} \in \mathcal{X}$
  - Feature vector $\mathrm{x} = (x_1, x_2, \dots, x_D)$
  - Feature space $\mathcal{X} = \mathbb{R}^D$
    - Think of D=1 in rest of slides, but Bayesian decision theory (Bayes classifier) holds for any D.
- Target/response $t \in \mathcal{Y}$
  - Discrete: Target space $\mathcal{Y} = \{C_1, C_2, \dots, C_K\}$
    - Often times also referred to as {1,2,..,K}, or for binary (K=2) classifiers as {0,1} or {-1,+1}
  - Continuous: Target space $\mathcal{Y} = \mathbb{R}$

- Classifier or regressor is simply a function from feature to target space
  - i.e., it maps each point in the feature space to a unique point in the target space
  - $h: \mathcal{X} \to \{C_1, \dots, C_K\}$
  - $\mathrm{y}: \mathcal{X} \to \mathbb{R}$

Notations (Bayes rule)

- $P(t|\mathrm{x}) = \frac{P(t)P(\mathrm{x}|t)}{P(\mathrm{x})} \propto P(t)P(\mathrm{x}|t)$

(posterior = prior x likelihood (class conditional) / evidence)

- $P(\mathrm{x}, t) \;=\; P(\mathrm{x})\,P(t|\mathrm{x}) \;=\; P(t)\,P(\mathrm{x}|t)$

(joint = evidence x posterior = prior x liklhd. (class cond.))

- For binary t, $P(\mathrm{x}) = P(t = C_1)P(\mathrm{x}|C_1) + p(t = C_2)P(\mathrm{x}|\,C_2)$

$$= P(C_1)P(\mathrm{x}|C_1) + P(C_2)P(\mathrm{x}|C_2)$$

Outline of Module M2

- M2. Bayesian Decision Theory (incl. Bayes classifiers)
    - M2.0 Introduction/Background (on Probability Theory)
    - **M2.1 Bayesian Decision Theory**
        - M2.1.0 Decision Theory for Classification/Regression (common defns./notations)
        - **M2.1.1 Decision Theory for Classification (Bayes classifiers)**
        - M2.1.2 Decision Theory for Regression (Squared loss, etc.)

M2.1.0 Decision Theory for Classification

- Inference step
- Determine either $\mathrm{p}(\mathrm{x}, \mathrm{t})$ or $p(t = C_k|\mathrm{x})$.

- Decision step
- For any given x, determine optimal class label $\mathrm{h}(\mathrm{x}) = \mathrm{C_j}$ for t.
- Optimality wrt *risk* or *expected loss* (misclassification rate or general loss function/matrix for binary vs. multi-class classifiers)

[CMB]

Bayes classifier (two classes)

- $h(\mathrm{x}) = C_1$ if $P(C_1|\mathrm{x}) > P(C_2|\mathrm{x})$
  $= C_2$ o.w (otherwise i.e., $P(C_2|\mathrm{x}) \geq P(C_1|\mathrm{x})$)

(Note: $P(C_1|\mathrm{x}) > P(C_2|\mathrm{x}) \Leftrightarrow$ ← for discriminative models
$P(C_1, \mathrm{x}) > P(C_2, \mathrm{x}) \Leftrightarrow$ ← for generative models
$\mathrm{P}(C_1)P(\mathrm{x}|C_1) > P(C_2)P(\mathrm{x}|C_2)$ ← for gen. models' learning)

- Bayes classifier is the ***optimal*** classifier among all classifiers
  - wrt minimizing the probability of error (aka misclassification rate), …
  - …assuming complete knowledge of the posterior distribution.

Optimality – minimum misclassification rate

- Let Decision region $R_i \coloneqq \{\mathrm{x} \in \mathcal{X} \mid h(\mathrm{x}) = C_i\}$

Optimality – minimum misclassification rate

- Let Decision region $R_i := \{\mathrm{x} \in \mathcal{X} \mid h(\mathrm{x}) = C_i\}$

$$P(\text{error}) = P_{x,t}(h(x) \neq t)$$

Optimality – minimum misclassification rate

- Let Decision region $R_i \coloneqq \{\mathrm{x} \in \mathcal{X} \mid h(\mathrm{x}) = C_i\}$

$$P(\text{error}) = P_{x,t}(h(x) \neq t)$$

$$= \int_x \sum_{t=C_1,C_2} p(x,t)\, \mathbb{1}_{\{h(x)\neq t\}}\, dx$$

Optimality – minimum misclassification rate

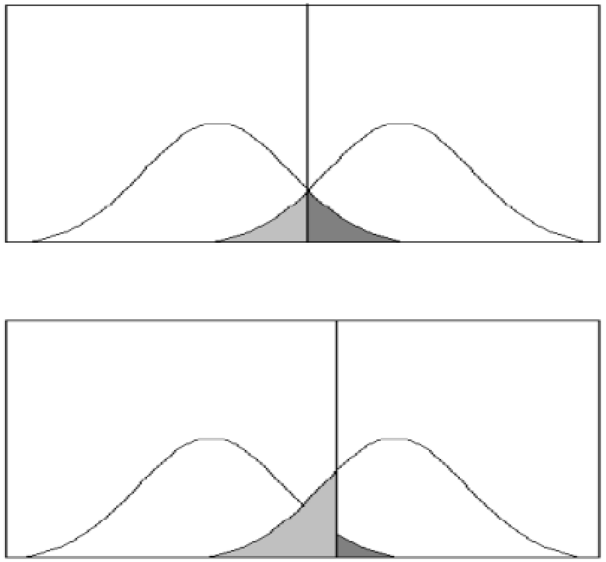
- Let Decision region $R_i := \{x \in \mathcal{X} \mid h(x) = C_i\}$

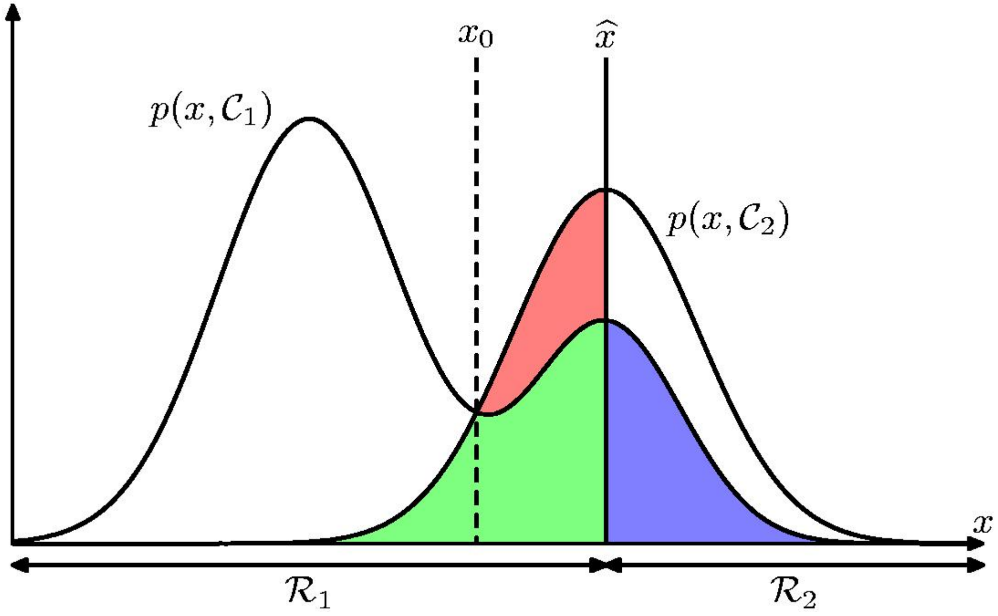
$$P(\text{error}) = P_{x,t}(h(x) \neq t)$$

$$= \int_x \sum_{t=C_1,C_2} P(x,t)\, \mathbb{1}_{\{h(x)\neq t\}}\, dx$$

$$= \int_x \sum_t P(t|x)\, \mathbb{1}_{\{h(x)\neq t\}}\, P(x)\, dx$$

Optimality - minimum misclassification rate



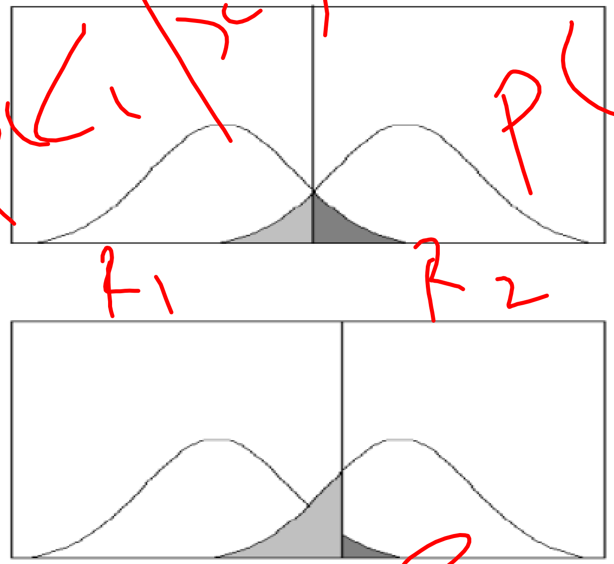


$$
\begin{aligned}
p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\
&= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\, \mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\, \mathrm{d}\mathbf{x}.
\end{aligned}
$$

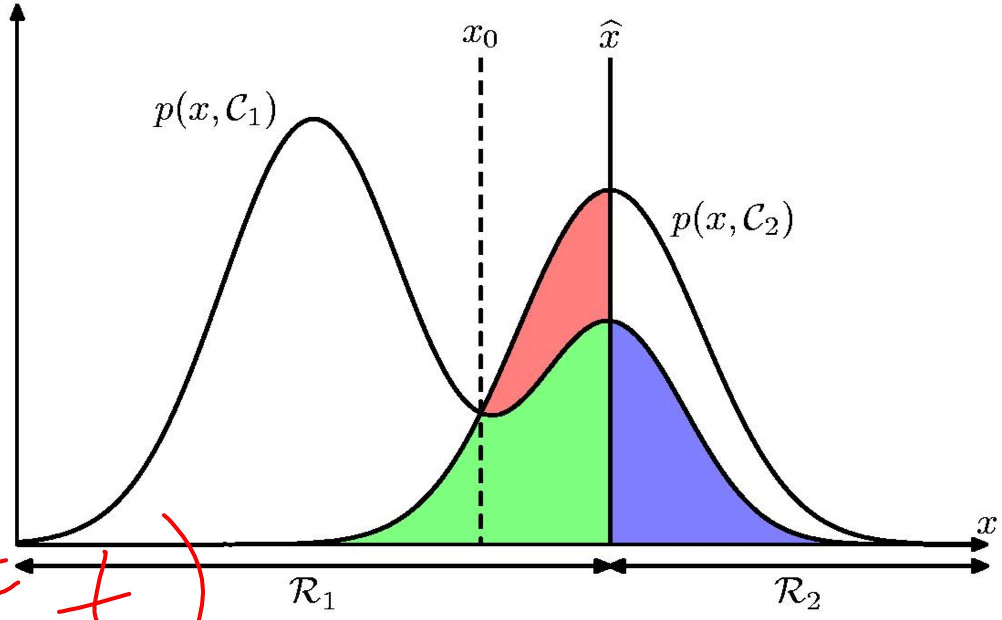
[CMB]

Optimality - minimum misclassification rate

$p(C_1|x)$ $p(C_2|x)$



R_1 R_2



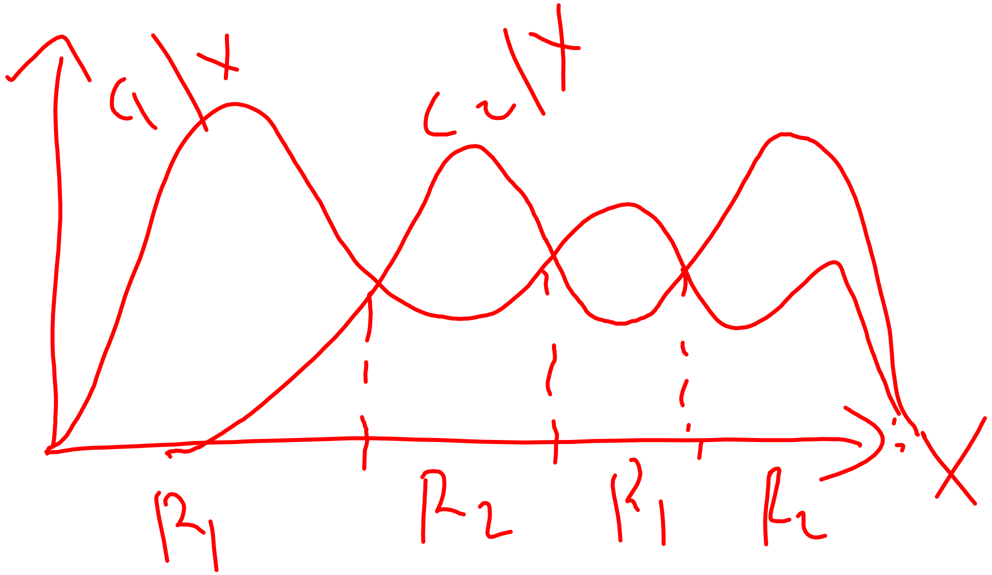
$= P_{x,t}(h(x) \neq t)$

$$
\begin{aligned}
p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\
&= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\,\mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\,\mathrm{d}\mathbf{x}.
\end{aligned}
$$

[CMB]

Can decision regions be discontiguous in the optimal classifier?

Can decision regions be discontiguous in the optimal classifier?



Bayes classifier (multi-class; K > 2 classes)

- $h(\mathrm{x}) \;=\; C_j \quad \text{if } P\left(t = C_j \,\middle|\, \mathrm{x}\right) \geq P\left(t = C_{j'} \,\middle|\, \mathrm{x}\right) \quad \forall j' \in \{1, \ldots, K\} \backslash \{\mathrm{j}\}$

  $\quad\quad\;\; = \operatorname{argmax}_{C_j} P\left(t = C_j \,\middle|\, \mathrm{x}\right) \quad (\text{ties broken arbitrarily})$

  $h(x) = C_1$ if $p(C_1|x) > p(C_2|x)$

- Again ***optimal*** classifier among all classifiers
  - wrt same criteria as for binary classifier i.e., minimum misclassification rate (or) equivalently maximum classification accuracy…
  - …assuming complete knowledge of the posterior distribution

Optimality of multi-class classifier (max. accuracy)

optimal h

$p(x_1)\,p(h(x_1)\,|\,x_1)$

$+\,p(x_2)\,p(h(x_2)\,|\,x_2)$

$$p(\text{correct}) = \sum_{k=1}^{K} p(\mathbf{x} \in \mathcal{R}_k, \mathcal{C}_k)$$

$$= \sum_{k=1}^{K} \int_{\mathcal{R}_k} p(\mathbf{x}, \mathcal{C}_k)\,\mathrm{d}\mathbf{x}$$

$h(x_1) = +1$

$h(x_2) = +1$

$h(x_3) = -1$

$h(x_4) = -1$

$p(+1\,|\,x_1) = 0.7$

$p(+1\,|\,x_3) = 0.2$

$p(+\,|\,x_2) = 0.4$

$p(+1\,|\,x_4) = 0.6$

Bayes classifier – General Loss Function (Matrix)

0-1 loss

True \ Pred	+1	-1	2	3
+1	0	1	1	1
-1	1	0	1	1
2	1	1	0	1
3	1	1	1	0

Bayes classifier – General Loss Function (Matrix)

- Example: classify medical images as ‘cancer’ or ‘normal’

Decision

$$\text{Truth}\quad\begin{array}{c c} & \begin{array}{cc} \text{cancer} & \text{normal} \end{array} \\ \begin{array}{c} \text{cancer} \\ \text{normal} \end{array} & \begin{pmatrix} 0 & 1000 \\ 1 & 0 \end{pmatrix} \end{array}$$

[CMB]

Optimality - Minimum Expected Loss

$$\mathbb{E}[L] = \sum_{t=C_1,\ldots,C_K} \sum_{j=C_1,\ldots,C_K} \int_{R_j} L_{tj}\, p(x,t) d\mathrm{x}$$

Regions $\mathcal{R}_j$ are chosen to minimize (next slides show why)

$$\mathbb{E}[L \mid X = \mathrm{x}\,] = \sum_{t=C_1,\ldots,C_K} L_{t,\,h(\mathrm{x})=j}\;\; p(t|\mathrm{x})$$

[CMB]

Optimality - Minimum Expected Loss

$$\mathbb{E}[L] = \sum_{t=C_1,\ldots,C_K} \sum_{j=C_1,\ldots,C_K} \int_{R_j} L_{tj}\, p(x,t)\,dx$$

Regions $\mathcal{R}_j$ are chosen to minimize (next slides show why)

$$\mathbb{E}[L \mid X = \mathrm{x}] = \sum_{t=C_1,\ldots,C_K} L_{t,\,h(\mathrm{x})=j} \;\; p(t|\mathrm{x})$$

1000 p(c₁|x)

1 × p(c₂|x)

⋮

L - 0-1 loss

$h(x) = \arg\max_{j} p(t=j|x)$

$\equiv \; h(x) = \arg\min_{j} \sum_{t} L_{tj}\, p(t|x)$

[CMB]

Optimality - Minimum Expected Loss (indicator fn. notation)

$$E[L] = \int_x \sum_{t=c_1}^{c_k} p(t|x) \left[\sum_{j=c_1}^{c_k} L_{tj} \mathbb{1}_{\{h(x)=j\}} \right] p(x)\, dx$$

$$= \int_x \sum_j \underbrace{\left(\sum_t L_{tj}\, p(t|x) \right)}_{\text{is minimized}} \mathbb{1}_{\{h(x)=j\}}\, p(x)\, dx$$

Choose $\underline{h(x)=j}$ s.t. $\underline{\hphantom{xx}}$ is minimized

Optimality - Minimum Expected Loss (cond. expectation notation)

$$\mathbb{E}_{x,t}[L] = \mathbb{E}_x[\mathbb{E}_{t|x}[L]] \quad (\mathbb{E}_x[\mathbb{E}_t[L \mid x]])$$

$$= \mathbb{E}_x[\mathbb{E}_{t|x}[L_{t,h(x)}]]$$

$$= \mathbb{E}_x\Big[\underbrace{\sum_{t=1}^{C} L_{t,h(x)}\, p(t|x)}\Big]$$

Choose $h(x) = j$ s.t. $\underline{\underline{\quad}}$ is minimized

[HR]

Cancer example – one final look!

- Example: classify medical images as ‘cancer’ or ‘normal’

Decision

$$\text{Truth}\quad \begin{array}{c c} & \begin{array}{cc} \text{cancer} & \text{normal} \end{array} \\ \begin{array}{c} \text{cancer} \\ \text{normal} \end{array} & \begin{pmatrix} 0 & 1000 \\ 1 & 0 \end{pmatrix} \end{array}$$

[CMB]

Cancer example – one final look!

- Example: classify medical images as 'cancer' or 'normal'

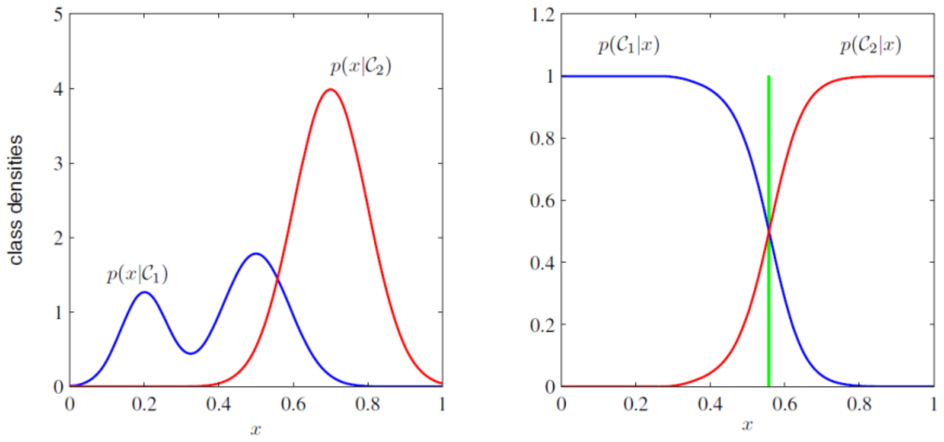
$$
\begin{array}{ccc}
 & & \text{Decision} \\
 & & \begin{array}{cc} \text{cancer} & \text{normal} \end{array} \\
\text{Truth} & \begin{array}{c} \text{cancer} \\ \text{normal} \end{array} & \begin{pmatrix} 0 & 1000 \\ 1 & 0 \end{pmatrix}
\end{array}
$$

$\underbrace{P(N|x)}_{\text{exp. loss if } h(x)=C}$ vs. $\underbrace{1000\,P(C|x)}_{\text{-do- if } h(x)=N}$

Inference and decision: three approaches for classification

- Generative model approach:
  (I) Model $p(x, C_k) = p(x|C_k)p(C_k)$
  (I) Use Bayes' theorem $p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})}$
  (D) Apply optimal decision criteria

- Discriminative model approach:
  (I) Model $p(\mathcal{C}_k|\mathbf{x})$ directly
  (D) Apply optimal decision criteria

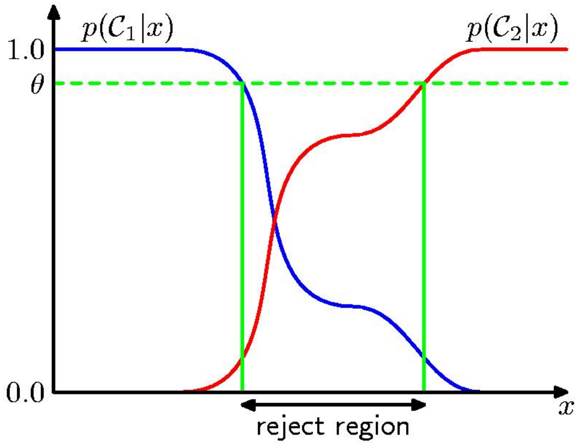


- Discriminant function approach:
  (D) Learn a function that maps each x to a class label directly from training data
  Note: No posterior probabilities!

[CMB]

Why separate Inference and Decision? (i.e., why infer (posterior) probabilities?)

- Minimizing risk (loss matrix may change over time)
- Reject option



- Combining models (Popular Naïve Bayes classifier)
- Etc.

[CMB]

Problem Setting

- Naïve Bayes Classifier
  - Assumption: The features are independent given the class labels

Problem Setting

- Naïve Bayes Classifier
  - Assumption: The features are independent given the class labels

Independent: $p(X_1, X_2) = p(X_1)\,p(X_2)$

$\not\Uparrow$

Conditionally independent: $p(X_1, X_2 | \underline{Y}) = p(X_1 | Y)\,p(X_2 | Y)$

Naïve Bayes

Naive Bayes assumption:

$$
\begin{aligned}
p(X|Y) &= p(X_1, X_2, \cdots, X_p|Y) \\
&= p(X_p|X_1, X_2, \cdots, X_{p-1}, Y)\, p(X_{p-1}|X_1, X_2, \cdots, X_{p-2}, Y) \cdots p(X_1|Y) \\
&= p(X_p|Y)\, p(X_{p-1}|Y) \cdots p(X_1|Y)
\end{aligned}
$$

$$
\begin{aligned}
p(Y|X) &= \frac{p(X_p|Y)\, p(X_{p-1}|Y) \cdots p(X_1|Y)\, p(Y)}{p(X)} \\
&\propto p(X_p|Y)\, p(X_{p-1}|Y) \cdots p(X_1|Y)\, p(Y)
\end{aligned}
$$

Naïve Bayes

- Assumption: The features are independent given the class labels
- Simple form for the probability distribution
- Not necessarily linear hyperplane ☺.
- Typically estimate by counting co-occurrences of feature value with class label
    - Maximum likelihood estimate
- Surprisingly powerful, especially in data with many features
    - High dimensional spaces

Understanding Bayes Theorem

c_1 = Inj

c_2 = NI

Given the data of accident reports and status as injured or not injured of the person after the accident.

$P(c_1) = 6/14$

$P(c_2) = 8/14$

c_1

$p(Y|c_1) = 1/2$

$p(NR|c_1) = 2/3$

c_2

$p(Y|c_2) = 1/8$

$p(NR|c_2) = 1/4$

$p(c_1|Y,NR) \propto p(Y|c_1)\, p(NR|c_1)\, p(c_1) = 1/7 = 0.88$

$p(c_2|Y,NR) \propto p(Y|c_2)\, p(NR|c_2)\, p(c_2) = 1/56 = 0.12$

Bike name	Repaired	Injured or Not injured
Yamaha	Yes	Injured
Yamaha	No	Injured
Suzuki	No	Not injured
TVS	Yes	Not injured
Honda	Yes	Not injured
Suzuki	Yes	Not injured
TVS	Yes	Injured
TVS	No	Injured
Honda	Yes	Not injured
Yamaha	No	Injured
Suzuki	Yes	Not injured
TVS	No	Injured
Honda	Yes	Not injured
Yamaha	No	Not injured

Case: Yamaha and Not repaired

Classification through Bayes Theorem

Given data on bikes and their features

$C_1 = Y$

$C_2 = H$

Bikes	weight	Engine
yamaha	100	300
yamaha	110	250
yamaha	92	250
yamaha	80	200
Honda	90	250
Honda	65	200
Honda	80	150
Honda	70	175

$p(85 \mid H)$

$p(W \mid H) \propto N(\mu, \sigma)$

$p(E \mid H)$

Predict the bike that was purchased from a given set of features,

Weight = 85 and engine = 250, Bike = ??

Where p(yamaha) = 0.5 and p(Honda) = 0.5

Assumptions

- Weight and engine are continuous variables
- Weight and engine are independent variables

Classification through Bayes Theorem

	Mean (weight)	Mean (Engine)	Variance (weights)	Variance (engine)
Yamaha(Y)	95.5	250	161	1666.66
Honda(H)	76.25	193.75	122.91	1822.91

Using Gaussian naïve Bayes,

P(Y/x(weight, engine)) = p(Y) *p(weight/Y)*p(engine/Y)*(1/p(x))

P(H/x) = p(H) *p(weight/H)*p(engine/H)*(1/p(x))

Using Gaussian distribution,

Probability	weight	engine
Yamaha	0.022331	0.009775
Honda	0.026361	0.003924

$$p(x=v|c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{-\frac{(v-\mu_c)^2}{2\sigma_c^2}}$$

P(yamaha/x) > p(Honda/x) $\Rightarrow$ YAMAHA