M2. Decision Theory (incl. Bayes classifiers)

Manikandan Narayanan

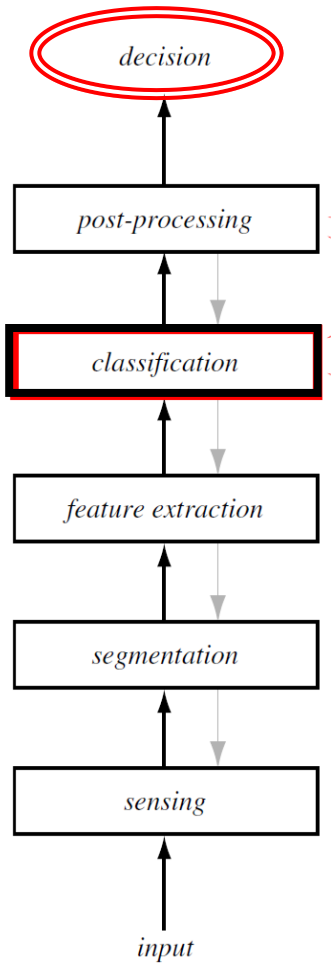
Week 2 (Aug 4-)

PRML Jul-Nov 2025 (Grads Section)

Recall: Full PRML pipeline

- Before we delve into the ML parts, let's also look at decision/action, the final step!!

- Expected learning outcomes of this topic:
  - ***primary:*** Understand Decision Theory
    - Optimal Bayes classifier
    - Optimal regressor
  - ***secondary:*** Understand certain paradigms/terms in ML:
    - Density estimation (discriminative vs. generative modelling) in the context of supervised learning (classification/regression), and
    - use it to set the stage for unsupervised learning topics like clustering and supervised topics like *Naïve* Bayes classifier!



Acknowledgment of Sources

- Slides based on content from related
  - Courses:
    - IITM – Profs. Arun/Harish[HR]/Chandra[CC]/Prashanth’s PRML offerings (slides, quizzes, notes, etc.), Prof. Ravi’s “Intro to ML” slides – cited (e.g., [HR]/[HG]) 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]
    - Foundations of ML by Mohri, Rostamizadeh, and Talwalkar (content, figures, slides by Mohri, etc.). – [MRT]

Outline of Module M2

- M2. Decision Theory (incl. Bayes classifiers)
  - **M2.0 Decision Theory for Classification/Regression (common defns./notations)**
  - M2.1 Decision Theory for Classification (Bayes classifiers)
  - M2.2 Decision Theory for Regression (Squared loss, etc.)

M2.0 Decision Theory (for classification/regression)

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

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

- Decision step
- For any given x, determine optimal t.
- Optimality wrt *(empirical) 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

$$\int_x f(x)\,dx \qquad f(x_1, \dots, x_D)$$
$$= \int_{x_1}\int_{x_2}\int_{x_D} f(x_1, \dots)\,dx_1 \dots dx_D$$

- 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
  - $h: \mathcal{X} \to \{C_1, \dots, C_K\}$
  - $f: \mathcal{X} \to \mathbb{R}$

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. Decision Theory (incl. Bayes classifiers)
  - M2.0 Decision Theory for Classification/Regression (common defns./notations)
  - **M2.1 Decision Theory for Classification (Bayes classifiers)**
  - M2.2 Decision Theory for Regression (Squared loss, etc.)

M2.1 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]

STOP & THINK: What is your guess for the optimal classifier (for binary classification)?

- That is, you are given a particular datapoint x.
- You already know $p(t = C_1|\mathrm{x})$ and $p(t = C_2|\mathrm{x})$ (say 0.3 and 0.7 respectively). Using this information,
  - how will you decide the optimal class label t for x?
    - Will your prediction be $\mathrm{h}(\mathrm{x}) = \mathrm{C}_1$ or $\mathrm{h}(\mathrm{x}) = \mathrm{C}_2$?
    - What is the expected 0-1 loss for each of these two cases?

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 proof – minimum misclassification rate -- in equations

$$E[L] = P(\text{error}) = P_{x,t}\left(h(x) \neq t\right)$$

$$= \int_x \sum_{t=c_1,c_2} P(x,t)\, \mathbb{1}_{\{h(x)\neq t\}}\, dx$$

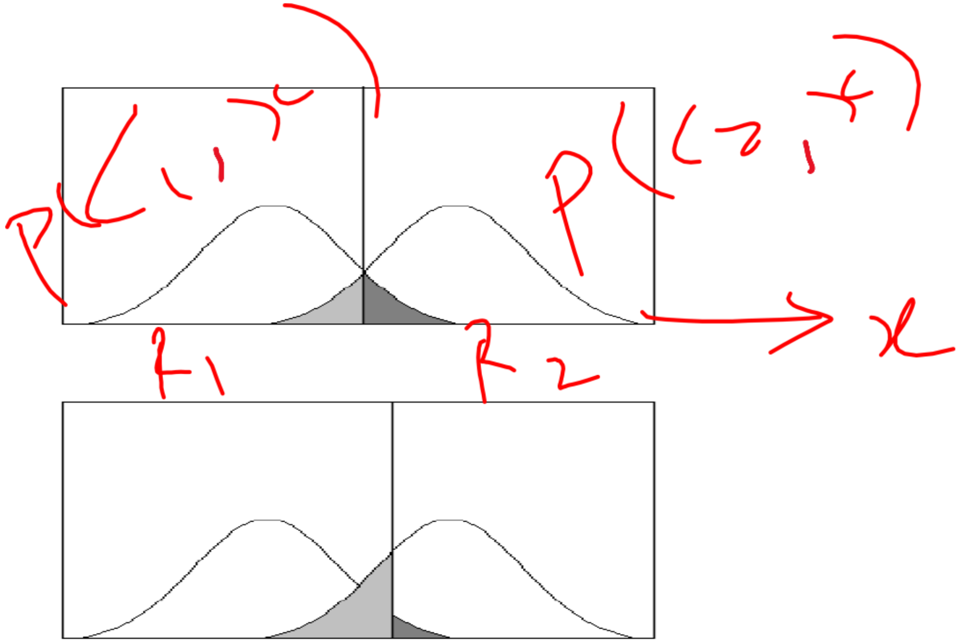
$$= \int_x \underbrace{\sum_t P(t|x)\, \mathbb{1}_{\{h(x)\neq t\}}}_{\text{min. by Bayes classifier in prev. slide}}\, P(x)\, dx$$

Optimality proof – minimum misclassification rate – in pictures

- **Goal:** Find $h: \mathcal{X} \rightarrow \{C_1, \ldots, C_K\}$ s.t. $R(h) = E[L(h)]$ is minimized.
- That is, find optimal classifier $h^* = \arg\min_h R(h)$

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

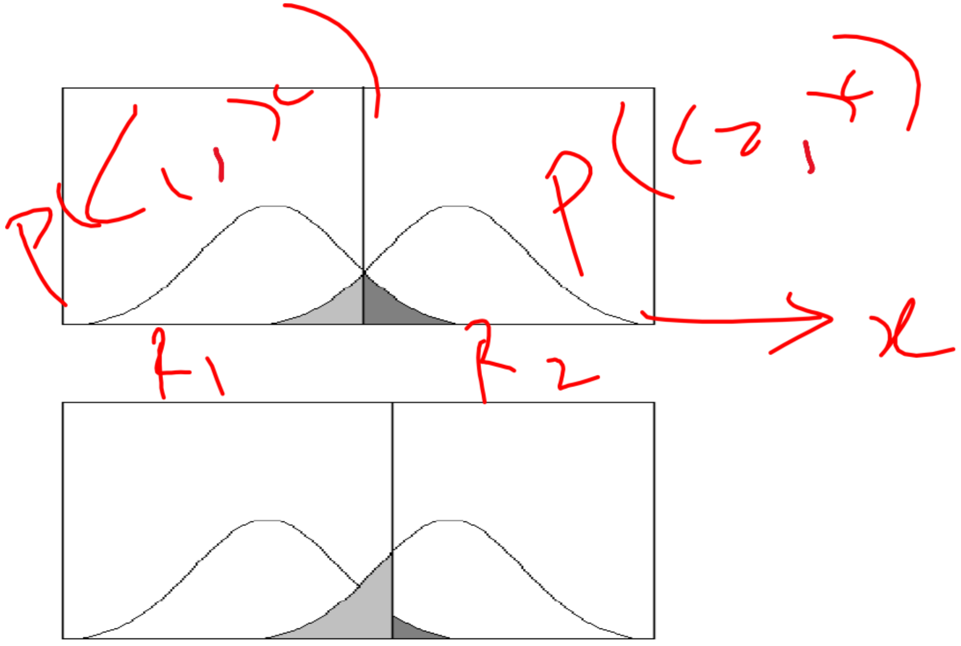
Optimality - minimum misclassification rate

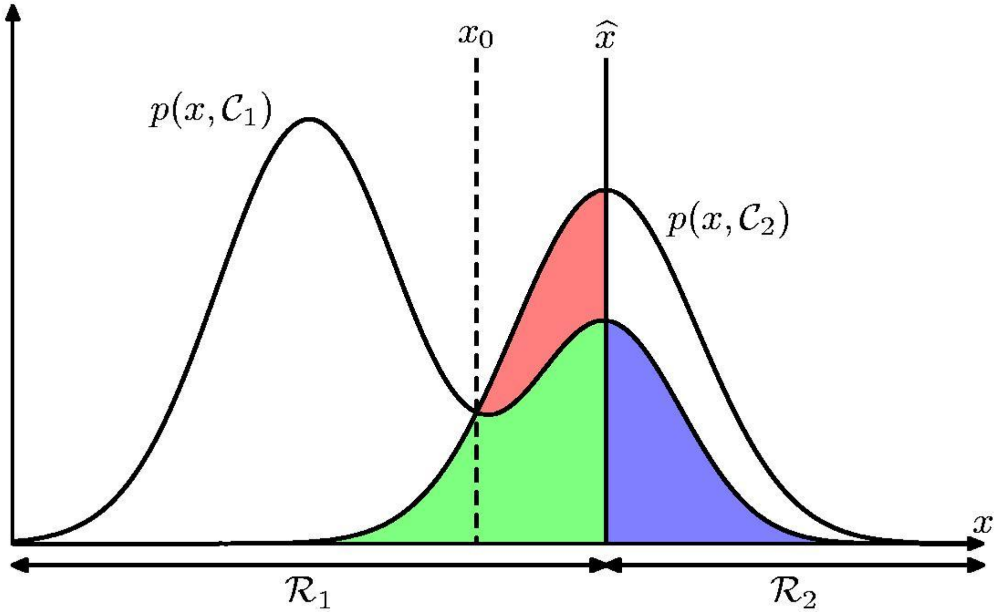


small note: $\boldsymbol{P}(\boldsymbol{t},\mathbf{x}) = P(t\,|\mathrm{x})\,P(\mathrm{x}) \propto \boldsymbol{P}(\boldsymbol{t}|\mathbf{x})$

[CMB]

Optimality - minimum misclassification rate



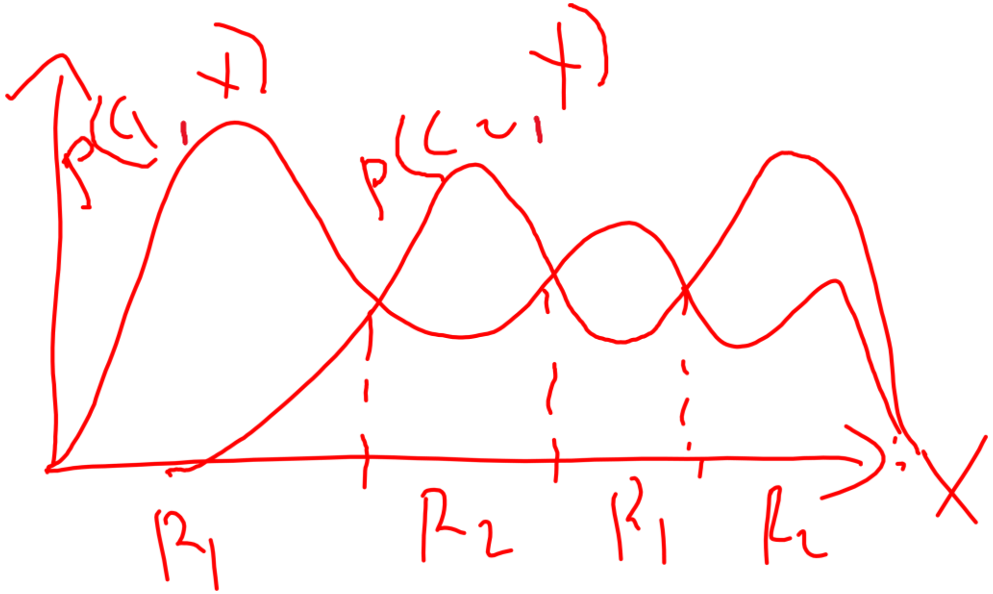


small note: $\boldsymbol{P}(\boldsymbol{t}, \mathbf{x}) = P(t \mid \mathrm{x})\, P(\mathrm{x}) \propto \boldsymbol{P}(\boldsymbol{t} \mid \mathbf{x})$

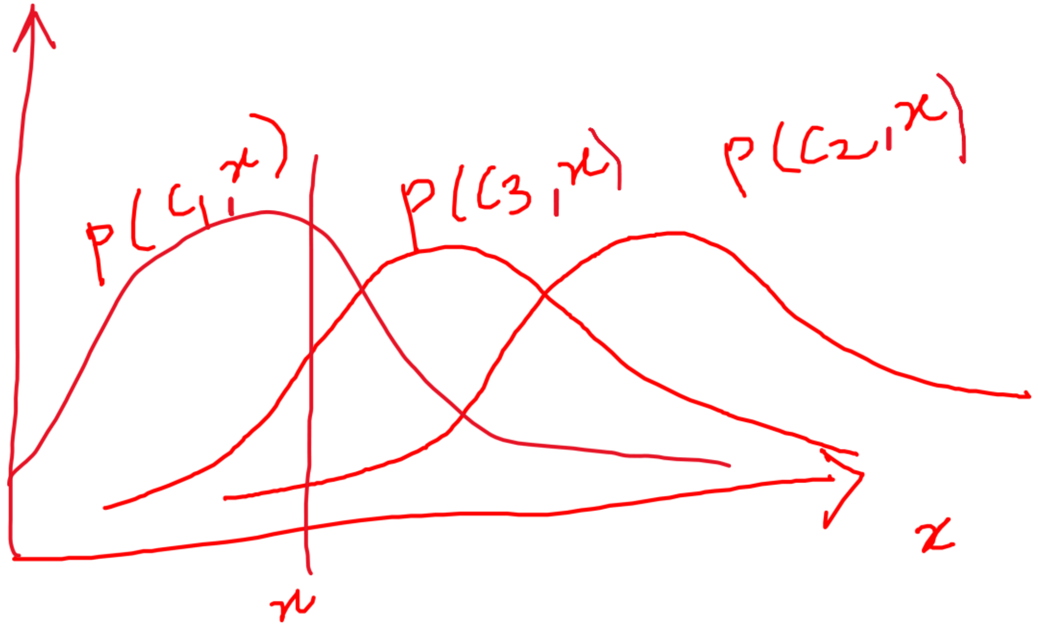
[CMB]

Can decision regions be discontiguous in the optimal classifier?

Can decision regions be discontiguous in the optimal classifier?



What about K > 2 classes?



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

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

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

- 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 – minimum misclassification rate -- in equations (for K > 2 classes)

$$E[L] = P(\text{error}) = \underset{x,t}{P}\left(h(x) \neq t \right)$$

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

$$= \int_x \underbrace{\sum_t P(t|x)\, \mathbb{1}_{\{h(x)\neq t\}}}_{\text{<min. by Bayes classifier in prev. slide>}} P(x)\, dx$$

Stop and Think! What have we seen so far?

- Optimizing mis-classification rate in K=2 and K > 2 settings.
- How about optimizing expected loss for general loss functions?
  - From 0—1 loss matrix to general loss matrix!

Bayes classifier – General Loss Function (Matrix)

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

$$\begin{array}{ccc} & \multicolumn{2}{c}{\text{Decision } h(x)} \\ & \text{cancer} & \text{normal} \end{array}$$

$$\begin{array}{rl} t \;\; \text{Truth} & \begin{array}{l} \text{cancer} \\ \text{normal} \end{array} \begin{pmatrix} 0 & 1000 \\ 1 & 0 \end{pmatrix} \end{array}$$

[CMB]

Bayes classifier – General Loss Function (Matrix)

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

Decision $h(X)$

y

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

<u>Intuition:</u>

$$h(x) = \begin{cases} N(\text{ormal}) & \text{if } P(N|x) > 1000 P(C|x) \\ C(\text{ancer}) & \text{o.w.} \end{cases}$$

Bayes classifier – General Loss Function (Matrix)

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

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

Intuition:

$$h(x) = \begin{cases} N(\text{ormal}) \text{ if } P(N|x) > 1000\, P(C|x) \\ C(\text{ancer}) \text{ o.w.} \end{cases}$$

$$
\begin{array}{cc}
 & (h(x)) \\
 & \begin{array}{ccccc} & 1 & 2 & \cdots & K \end{array} \\
L = (t) & \begin{array}{c} 1 \\ 2 \\ \vdots \\ K \end{array} \begin{bmatrix} 0 & & & 100 \\ & 0 & & \\ & L_{tj} & \ddots & \\ & & & 0 \end{bmatrix}
\end{array}
$$

[CMB]

Optimality - Minimum Expected Loss (integration notation)

$$\mathbb{E}_{x,t}[L] = \int_x \sum_t L_{t,h(x)}\, p(x,t)\, dx$$

$$= \int_x \underbrace{\sum_{t=C_1}^{C_K} p(t|x)\, L_{t,h(x)}}_{①}\, p(x)\, dx$$

Choose $\underline{h(x)=j}$ s.t. ① is minimized.

$$E_{t|X}[L_{t,h(X)}]$$

Optimality - Minimum Expected Loss (contd. expectation notation)

$$\mathbb{E}_{x,t}[L] = \mathbb{E}_x\left[\mathbb{E}_{t|x}[L]\right]$$

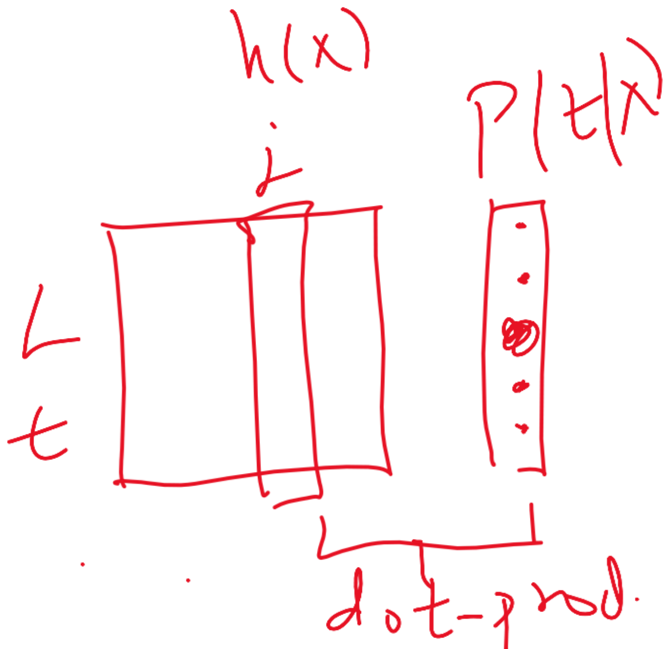
$$= \mathbb{E}_x\left[\mathbb{E}_{t|x}\left[L_{t,h(x)}\right]\right]$$

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

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

[HR]

Optimality - Minimum Expected Loss – Minimize $\mathbf{E}_{\mathrm{t}|\mathrm{x}}[L_{t,j}]$ (aka dot-prod.) over all j.



$$h(x) = \arg\min_{j} \sum_{t} L_{tj} \cdot P(t|x)$$

[CMB]

Cancer example – one final look!

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

$$
\begin{array}{c c c c}
 & & \multicolumn{2}{c}{\text{Decision}} \\
 & & \text{cancer} & \text{normal} \\
\text{Truth} & \begin{array}{l}\text{cancer}\\ \text{normal}\end{array} & \multicolumn{2}{c}{\begin{pmatrix} \boxed{\begin{matrix}0\\1\end{matrix}} & \begin{matrix}1000\\0\end{matrix}\end{pmatrix}}
\end{array}
\qquad P(t|x)
$$

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

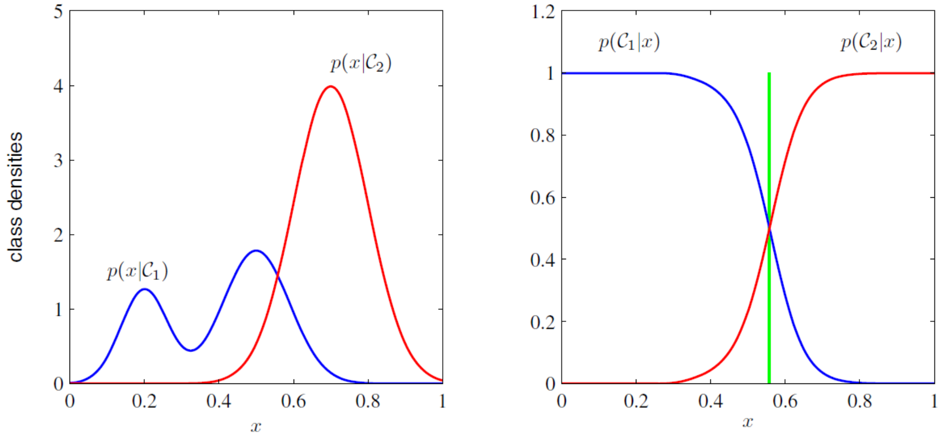
matches Intuition:

$$h(x) = \begin{cases} N(\text{ormal}) & \text{if } P(N|x) > 1000\,P(C|x) \\ C(\text{ancer}) & \text{o.w.} \end{cases}$$

[CMB]

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



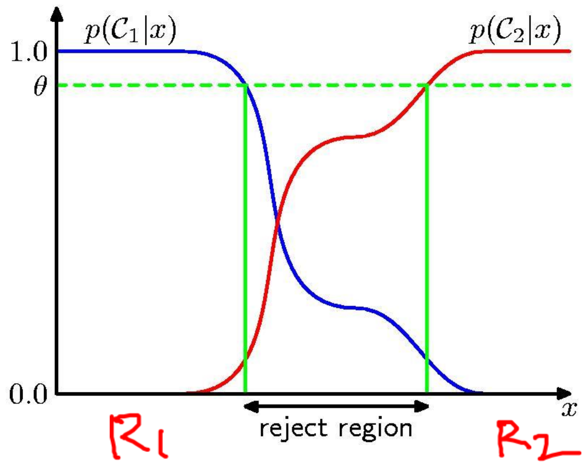
- Discriminative model approach:
  - (I) Model $p(\mathcal{C}_k|\mathbf{x})$ directly
  - (D) Apply optimal decision criteria

- 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]

Example of generative vs. discriminative models for the same task

- Task: Classification
  - Discriminative model p(t|x) : Logistic regression
  - Generative model p(t,x) = p(t) p(x|t) : Naïve Bayes classifier

- In general, discriminative model preferred – folklore
  - But a nuance: discriminative preferred for large sample sizes, vs. generative for smaller sample sizes (IF model assumptions are satisfied!)

    [Optional reference: On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes. *Andrew Ng* and *Michael Jordan. NIPS 2021.*]

Final notes for tutorial

- **Exercise:** What is Precision, Recall, etc., in terms of Probab. over (x, t)?

- **Loss matrix $\neq$ Confusion matrix,**
  - but to estimate R[h]=E[L(h)] of a learned (trained) classifier h(.), we can use both matrices.
  - ***Important:*** To estimate R[h], use confusion matrix wrt **test data**. Why??

[HR]

Some examples

- See Bayes classifier examples in [HG]Notes!

[HR]

Outline of Module M2

- M2. Decision Theory (incl. Bayes classifiers)
  - M2.0 Decision Theory for Classification/Regression (common defns./notations)
  - M2.1 Decision Theory for Classification (Bayes classifiers)
  - **M2.2 Decision Theory for Regression (Squared loss, etc.)**

M2.2 Decision Theory for Regression

- Inference step
- Determine $p(\mathbf{x}, t)$ or $p(t \mid \boldsymbol{x})$.

- Decision step
- For given x, make optimal prediction f(x) for t (min. risk or expected loss).
- Given a loss fn., $E[L] = \iint L\big(t, f(\boldsymbol{x})\big)\, p(\boldsymbol{x}, t)\, d\boldsymbol{x}\, dt$

[CMB]

The Squared Loss Function

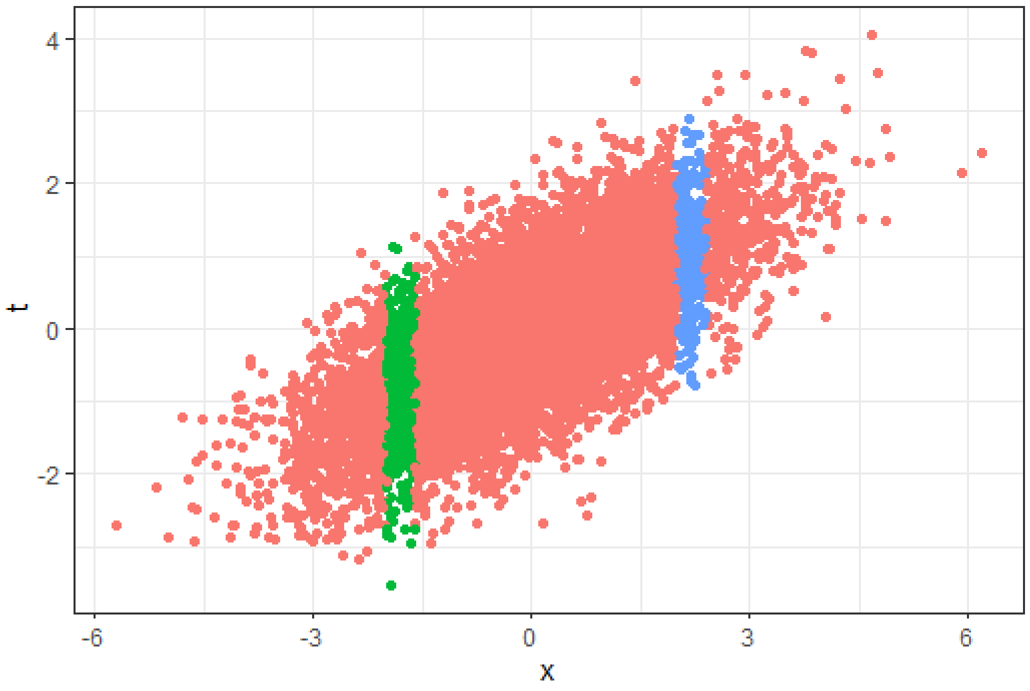
$$E[L] = \iint (f(\boldsymbol{x}) - \boldsymbol{t})^2 \, p(\boldsymbol{x}, t) \, d\boldsymbol{x} \, dt$$

$$E[L] = \int_x \int_t (f(x) - t)^2 p(t|x) dt \; p(x) dx$$

$$E_{X,t}[L] = E_x \left[E_{t|x} \left[(f(x) - t)^2 \right] \right]$$

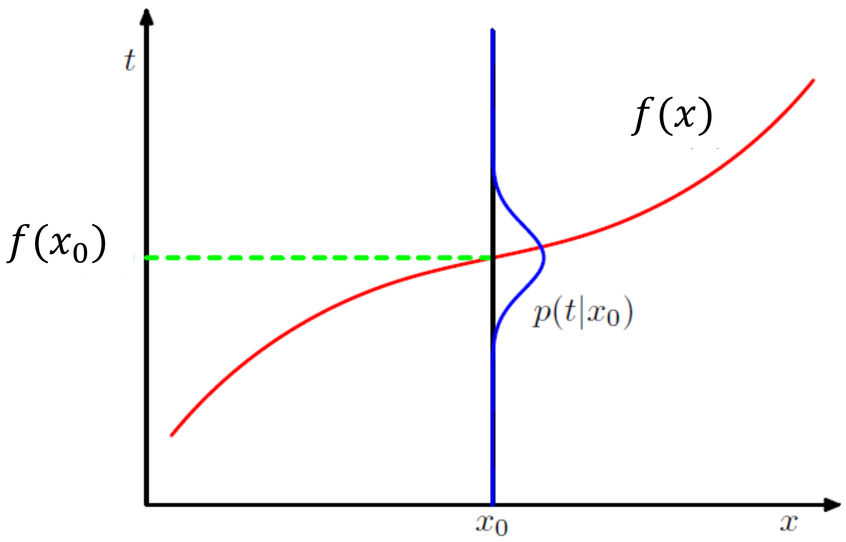
[CMB]

What would be a good minimizer if (x,t) is bivariate normally distributed like in our previous example ($\boldsymbol{X} = W_1 + W_2,\ t = W_2$)? (with W_1, W_2 indept. std. Gaussian rvs.)



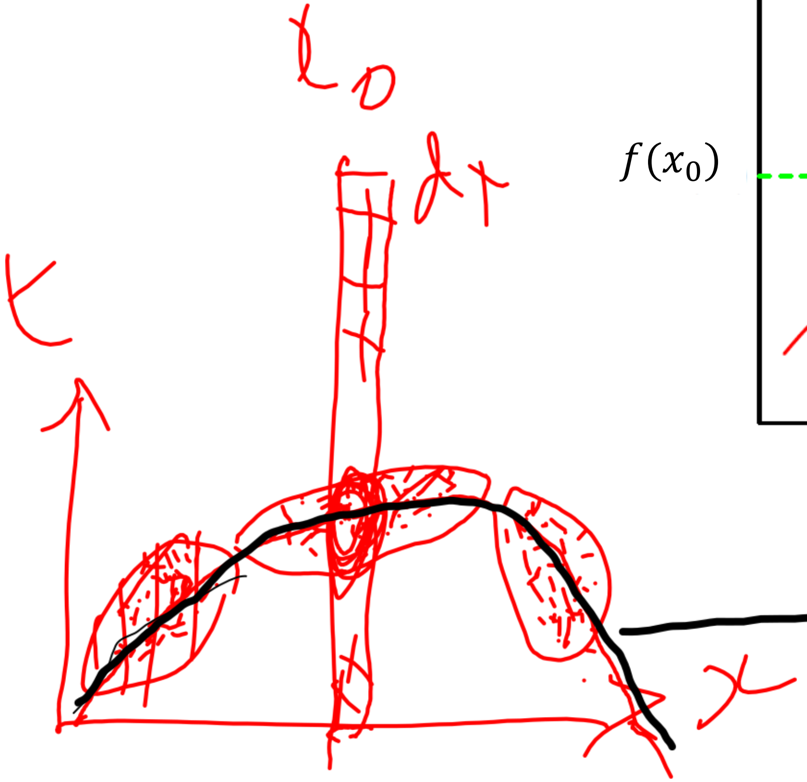
[HR]

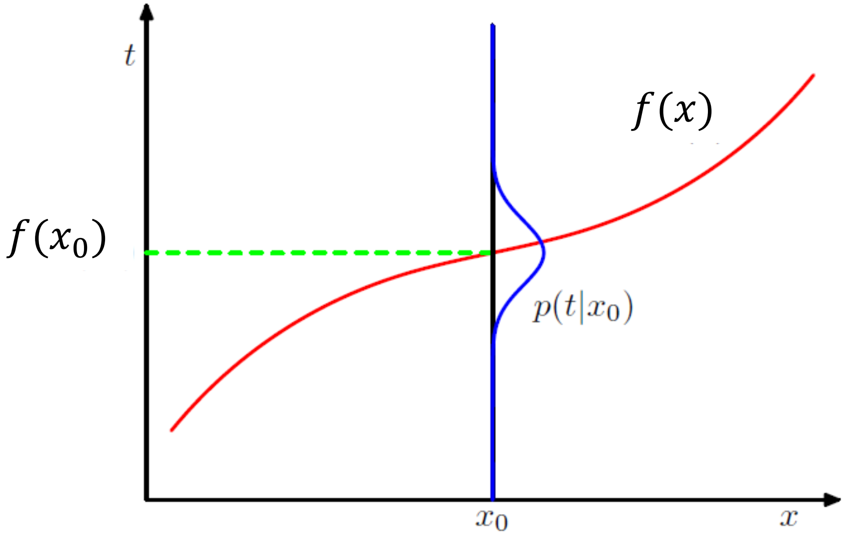
What if (x,t) has some general distribution?



[CMB]

What if (x,t) has some general distribution?





$$t \quad \underbrace{\int p(t|x)\, t\, dt}_{E[t|x]}$$

[CMB]

Optimality – min. squared loss (conditional expectation as a minimizer)

$$E[L] = E_x\left[E_{t|x}\left[(f(x)-t)^2\right]\right]$$

$$\{f(x)-t\}^2 = \{f(x) - E[t|x] + E[t|x] \ - t\}^2$$

$$= \{f(x) - E[t|x]\}^2 + 2\{f(x) - E[t|x]\}\{E[t|x] - t\} + \{E[t|x] - t\}^2$$

$$E_{t|x}\left[(f(x)-t)^2\right] =$$

[CMB]

Optimality – min. squared loss (conditional expectation as a minimizer)

$E[L] = E_x\left[E_{t|x}\left[(f(x)-t)^2\right]\right]$

$$\{f(x) - t\}^2 = \{f(x) - E[t|x] + E[t|x] \; - t\}^2$$

$$= \{f(x) - E[t|x]\}^2 + \underbrace{2\{f(x) - E[t|x]\}\{E[t|x] - t\}}_{0} + \underbrace{\{E[t|x] - t\}^2}_{\boxed{Var(t|x)}}$$

$E_{t|x}\left[(f(x)-t)^2\right] = \ldots$

$$\mathrm{E[L]} = \int \{f(x) - E[t|x]\}^2 \, p(x)dx + \underbrace{\int var(t|x)p(x)dx}_{\text{noise}}$$

$$y(x) = E[t|x]$$

[CMB]

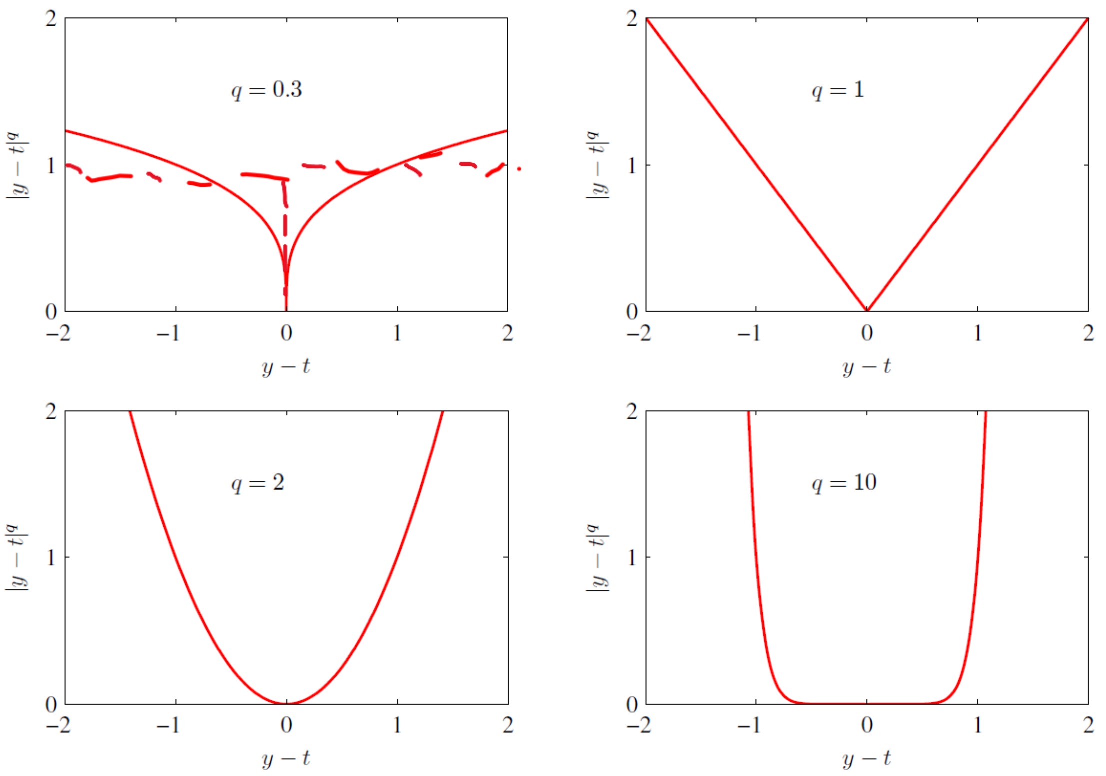
From Squared to Minkowski loss

$$E[L_q] = \iint |f(\boldsymbol{x}) - t|^q \, p(\boldsymbol{x}, t) \, d\boldsymbol{x} \, dt$$

Conditional mean / median / mode

for q = 2 / q=1 / q→0

respec.



<u>p(t|x) – inherent variab. in data</u>

[CMB]

Three approaches again (for regression)

- Generative model approach:

  (I) Model $p(t, \mathbf{x}) = p(\mathbf{x}|t)p(t)$

  (I) Use Bayes' theorem $p(t|\mathbf{x}) = \dfrac{p(\mathbf{x}|t)p(t)}{p(\mathbf{x})}$

  (D) Take conditional mean/median/mode/any other optimal decision outcome as f(x)

- Discriminative model approach:

  (I) Model $p(t|\mathbf{x})$ directly

  (D) Take conditional mean/median/mode/any other optimal decision outcome as f(x)

- Direct regression approach:

  (D) Learn a regression function f(x) directly from training data

[CMB]

In summary

- Decision theory (for classifn./regn.):
  - inference and decision steps
    - Generative vs. discriminative models for inference
    - Minimum risk (expected loss) for optimal decision
    - Different loss functions possible for classification; similarly for regression; but require knowledge of posterior (or joint directly or via prior and liklhd.) densities
  - Direct/discriminant approach also possible to take decision

- Optimal models, but how to learn them?
  - Bayes classifier for classifn. and conditional mean for regn., PROVIDED we've access to the joint or posterior distbn.
  - Next → Q: How do we learn these distbns. (and hence the classifier/regressor) from data? A: Density estimation of P(X,Y) or P(Y|X).

Thank you!

Backup slides follow

Backup

Optimality of multi-class classifier (max. accuracy)

$$p(\text{error}) = P_{x,t}(h(x) \neq t) = 1 - \underbrace{P_{x,t}(h(x) = t)}_{P(\text{correct})}$$

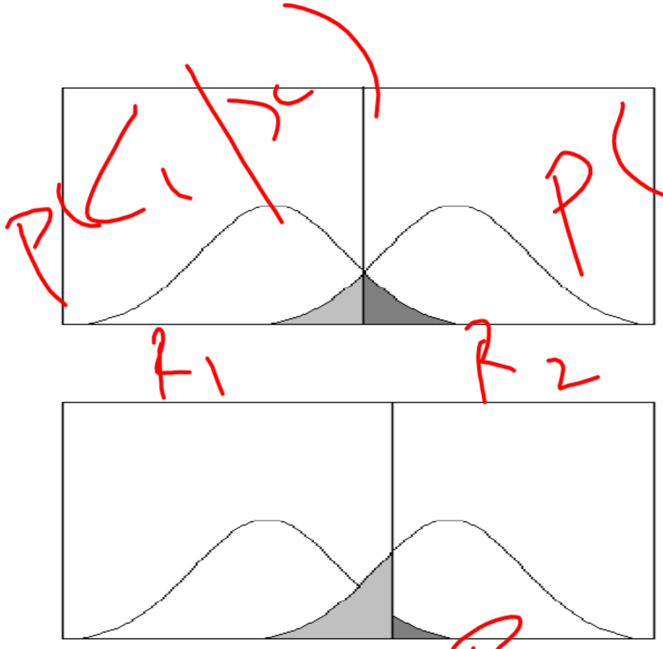
$$\mathbb{1}_c = \begin{cases} 1 & \text{if } c \text{ is T} \\ 0 & \text{if } c \text{ is F} \\ & \text{alse} \end{cases}$$

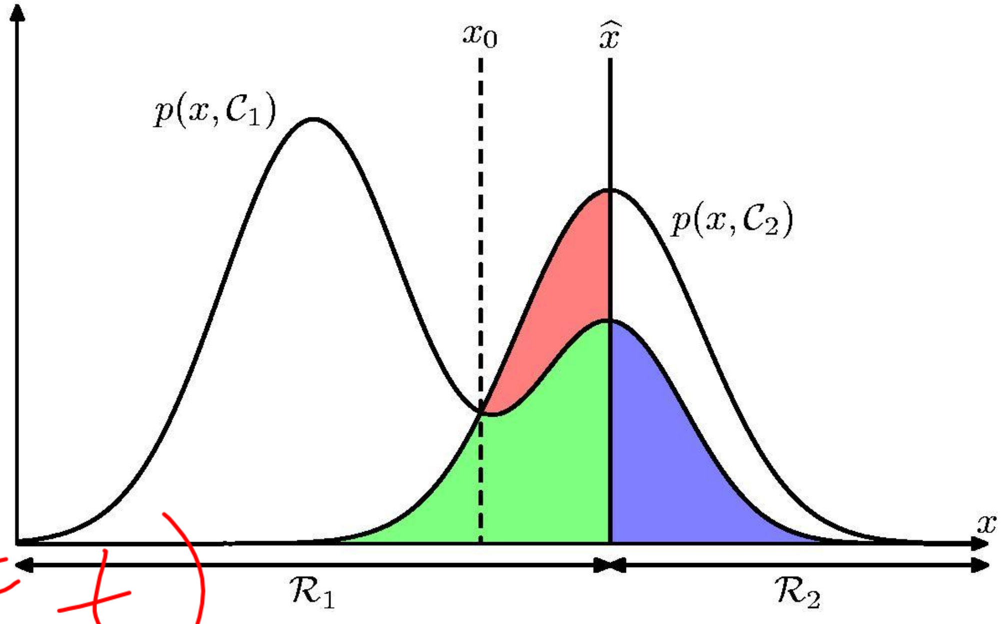
$P(\text{correct})$

$$\hookrightarrow \int_x \sum_{t=c_1}^{c_K} p(x,t)\, \mathbb{1}_{\{h(x)=t\}}\, dx = \int_x \left(\sum_{t=c_1}^{c_K} p(t|x)\, \mathbb{1}_{\{h(x)=t\}} \right) p(x)\, dx$$

[CMB]

Optimality - minimum misclassification rate





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

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

[CMB]

Optimality of multi-class classifier (min. error)

p(error)

$$= \left(\int_x \sum_{t=1}^{K} p(x,t) \, \mathbb{1}_{(h(x) \neq t)} \, dx \right)$$

$$p(x, \underset{=}{t} C_1) + p(x, \overset{t}{C_2}) + \cdots + p(x, \overset{t=}{C_K})$$

$$\left[p(x) - \underline{\underline{p(x, t=C_2)}} \right]$$

Optimality - Minimum Expected Loss (indicator fn. notation)

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

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

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