Lecture 3
Estimation Theory

- Formulating the problem
- Maximum Likelihood Estimation
- Maximum a posteriori Estimation
- Naive Bayes and logistic regression
- Minimum Mean-Square Error Estimation

Contents

- Formulating the problem
- Maximum Likelihood Estimation
- Maximum a posteriori Estimation
- Naive Bayes and logistic regression
- Minimum Mean-Square Error Estimation

What is optimal?

- The “goodness" of an estimate can be expressed in different ways, depending upon the particular engineering problem
- 3 commonly-used optimality criterion: the maximum-likelihood, maximum a posteriori, and minimum mean-square error criterion

Notation

$\mathbf{s}(n)$	signal	$s(n)$	signal realization
$\mathbf{v}(n)$	noise signal	$v(n)$	noise signal realization
$\mathbf{z}(n)$	sample	$z(n)$	sample realization
$\hat{\mathbf{s}}(n)$	estimate	$\hat{s}(n)$	a specific estimate
$\tilde{\mathbf{s}}(n) = \mathbf{s} - \hat{\mathbf{s}}$	estimation error	$\tilde{s}(n)$	a specific estimation error

Optimal Estimation Problem

Given the measurements $\mathbf{z}(1), \mathbf{z}(2), \ldots, \mathbf{z}(n)$, the corruption function g such that

$$\mathbf{z}(n) = g(\mathbf{s}(n), \mathbf{v}(n), n)$$

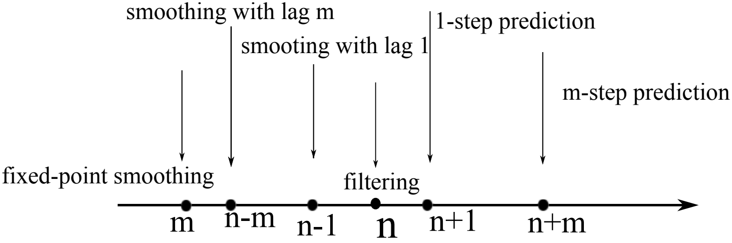
and an optimality criterion. Design an estimator that generates an optimal estimate $\hat{\mathbf{s}}(n)$ of $\mathbf{s}(n)$ given by

$$\hat{\mathbf{s}}(n) = \alpha_n(\mathbf{z}(1), \mathbf{z}(2), \ldots, \mathbf{z}(n)),$$

for some function α_n.

Prediction, Filtering and Smoothing

Given observations $\{\mathbf{z}(1), \mathbf{z}(2), \dots, \mathbf{z}(n)\}$, make the best guess of the value of $\mathbf{s}(\ell)$.



Properties of the Estimates

- Unbiased estimator:

$$E(\hat{\mathbf{s}}(n)) = E(\mathbf{s}(n)), E(\tilde{\mathbf{s}}(n)) = 0$$

- Asymptotically unbiased estimator:

$$\lim_{n\to\infty} E(\hat{\mathbf{s}}(n)) = E(\mathbf{s}(n)), \lim_{n\to\infty} E(\tilde{\mathbf{s}}(n)) = 0$$

- Consistent estimator:

$$\lim_{n\to\infty} E(\tilde{\mathbf{s}}^2(n)) = 0$$

Example: the mean filter

Constant signal $\mathbf{s}$: $\mathbf{s}(n) = \mathbf{s}$

measurement: $\mathbf{z}(n) = \mathbf{s} + \mathbf{v}(n)$

$\mathbf{v}(n)$ is a RV with mean 0 and variance σ_v^2

independence of $\mathbf{s}$ and $\mathbf{v}(1), \mathbf{v}(2), \ldots, \mathbf{v}(n)$

The mean filter

$$\hat{\mathbf{s}}(n) = \frac{1}{n} \sum_{j=1}^{n} \mathbf{z}(j)$$

Example

Check the properties of the estimate

- unbiasedness: $E[\hat{\mathbf{s}}(n)] = E[\mathbf{s}]$
- consistency: $E[\tilde{\mathbf{s}}^2(n)] = \frac{\sigma_v^2}{n}$

Contents

- Formulating the problem
- Maximum Likelihood Estimation
- Maximum a posteriori Estimation
- Naive Bayes and logistic regression
- Minimum Mean-Square Error Estimation

Formulating the maximum likelihood estimation problem

- most-probable (most likely) $\longrightarrow$ the peak of $f_{\mathbf{x}}(x)$, i.e., most-likely value of $\mathbf{x}$=value of x that maximizes $f_{\mathbf{x}}(x)$
- A single measurement z $\longrightarrow$ find the value of s that is **most likely** to have produced z $\longrightarrow$ seek the value of s that maximizes the **likelihood function**.
- **Likelihood function:** $f_{\mathbf{z}}(z|\mathbf{s} = s)$
- Maximum likelihood Estimation:

$$\hat{\mathbf{s}}_{\mathrm{ML}} = \text{value of } s \text{ that maximizes } f_{\mathbf{z}}(z|\mathbf{s} = s)$$

Derivation of the estimation

$$\hat{\mathbf{s}}_{\mathrm{ML}} = \text{value of } s \text{ for which } \frac{\partial f_{\mathbf{z}}(z|\mathbf{s}=s)}{\partial s} = 0$$

log-likelihood function

$$\hat{\mathbf{s}}_{\mathrm{ML}} = \text{value of } s \text{ for which } \frac{\partial \ln f_{\mathbf{z}}(z|\mathbf{s}=s)}{\partial s} = 0$$

likelihood function VS density?

Example

Suppose $\mathbf{s}$ and $\mathbf{z}$ are random variables with joint pdf

$$f_{\mathbf{s},\mathbf{z}}(s,z) = \begin{cases} \frac{1}{12}(s+z)e^{-z}, & 0 \leq s \leq 4, 0 \leq z \leq \infty; \\ 0, & \text{otherwise;} \end{cases}$$

The goal is to compute the ML estimate of $\mathbf{s}$ based on z.

Example

Find the likelihood function:

$$f_{\mathbf{z}}(z|\mathbf{s}=s) = \frac{f_{\mathbf{s},\mathbf{z}}(s,z)}{f_{\mathbf{s}}(s)}$$

As

$$f_{\mathbf{s}}(s) = \int_0^\infty f_{\mathbf{s},\mathbf{z}}(s,z)dz = \frac{1}{12}(s+1),\ 0 \le s \le 4.$$

the likelihood function is

$$f_{\mathbf{z}}(z|\mathbf{s}=s) = \frac{s+z}{s+1}e^{-z},\ 0 \le s \le 4, 0 \le z \le \infty$$

Example

Find the value of s that maximizes $f_{\mathbf{z}}(z|\mathrm{s}=s)$.

Calculate the partial derivative as:

$$\frac{\partial f_{\mathbf{z}}(z|\mathbf{s}=s)}{\partial s} = \frac{1-z}{(s+1)^2}e^{-z}$$

Hence

$$\hat{s}_{\mathrm{ML}} = \begin{cases} 4 & 0 \le z < 1 \\ 2 & z = 1 \\ 0 & z > 1 \end{cases}$$

Example: ML Estimation with Gaussian Noise

Suppose that $\mathbf{z} = \mathbf{s} + \mathbf{v}$, where $\mathbf{s}$ and $\mathbf{v}$ are independent and $\mathbf{v} \sim \mathcal{N}(0, \sigma^2)$, i.e.,

$$f_{\mathbf{v}}(v) = \frac{1}{\sqrt{2\pi}\sigma} e^{-v^2/2\sigma^2}$$

Given the sample realization z, derive the maximum likelihood estimation $\hat{s}_{\mathrm{ML}}$.

Example: ML Estimation with Gaussian Noise

As $F(z|\mathbf{s}=s)=P(\mathbf{z}\le z|\mathbf{s}=s)=\frac{P(\mathbf{z}\le z,\mathbf{s}=s)}{P(\mathbf{s}=s)}=P(\mathbf{v}\le z-s)$, the likelihood function is

$$f_{\mathbf{z}}(z|\mathbf{s}=s)=f_{\mathbf{v}}(v)|_{v=z-s}=\frac{1}{\sqrt{2\pi}\sigma}e^{-(z-s)^2/2\sigma^2}$$

Thus

$$\hat{s}_{\mathrm{ML}}=z,\hat{\mathbf{s}}_{\mathrm{ML}}=\mathbf{z}.$$

Contents

- Formulating the problem
- Maximum Likelihood Estimation
- Maximum a posteriori Estimation
- Naive Bayes and logistic regression
- Minimum Mean-Square Error Estimation

Maximum *a posteriori* Estimation

Another optimality criterion:

maximizes the conditional density $f_{\mathbf{s}}(s|\mathbf{z} = z)$

The density is known as the *a posteriori* density since it is the density after the measurement z becomes available.

the maximum a posteriori (MAP) estimate

$$\hat{\mathbf{s}}_{\mathrm{MAP}} = \text{value of } \mathbf{s} \text{ that maximizes } f_{\mathbf{s}}(s|\mathbf{z} = z)$$

Maximum *a posteriori* Estimation

Assuming $f_{\mathbf{s}}(s|\mathbf{z}=z)$ is differentiable and has a unique maximum in the interior of its domain, we have

$$\hat{\mathrm{s}}_{\mathrm{MAP}} = \text{value of } \mathrm{s} \text{ for which } \frac{\partial f_{\mathbf{s}}(s|\mathbf{z}=z)}{\partial s} = 0,$$

By Bayes' formula,

$$f_{\mathbf{s}}(s|\mathbf{z}=z) = \frac{f_{\mathbf{z}}(z|\mathrm{s}=s) f_{\mathbf{s}}(s)}{f_{\mathbf{z}}(z)}$$

Thus,

$$\hat{\mathrm{s}}_{\mathrm{MAP}} = \text{value of } \mathrm{s} \text{ that maximizes } f(z|\mathrm{s}=s) f_{\mathbf{s}}(s).$$

ML Estimate Vs MAP Estimate

- $\max_s f_{\mathbf{z}}(z|\mathbf{s}=s)$ Vs $\max_s f_{\mathbf{s}}(s|\mathbf{z}=z)$
  - In likelihood you have observed some outcome, so you want to find/create/estimate the most likely source/model/parameter/probability distribution from which this event has raised, i.e., likelihood attaches to hypotheses.
  - In probability you usually want to find the probability of a possible event based on a model/parameter/probability distribution, i.e., probability attaches to possible results.
- $\max_s f_{\mathbf{z}}(z|\mathbf{s}=s)$ Vs $\max_s f_{\mathbf{z}}(z|\mathbf{s}=s)f_{\mathbf{s}}(s)$
- In MAP estimate, the density $f_{\mathbf{s}}(s)$ must be known, i.e., it is a Bayesian estimation

Example: MAP estimation with Gaussian noise

Additive-noise $\mathbf{z} = \mathbf{s} + \mathbf{v}$, where $\mathbf{v} \sim \mathcal{N}(0, \sigma_v^2)$. Assume $\mathbf{s} \sim \mathcal{N}(\eta_s, \sigma_s^2)$.

Then

$$f_{\mathbf{s}}(s) = \frac{1}{\sqrt{2\pi}\sigma_s} e^{-\frac{(s-\eta_s)^2}{2\sigma_s^2}}$$

As

$$f_{\mathbf{z}}(z|\mathbf{s} = s) = \frac{1}{\sqrt{2\pi}\sigma_v} e^{-(z-s)^2/2\sigma_v^2}$$

we have

$$f(z|\mathbf{s} = s) f_{\mathbf{s}}(s) = \frac{1}{2\pi\sigma_s\sigma_v} \exp\left[-\frac{(z-s)^2}{2\sigma_v^2} - \frac{(s-\eta_s)^2}{2\sigma_s^2}\right]$$

Example: MAP estimation with Gaussian noise

Differentiating the term $\left[-\frac{(z-s)^2}{2\sigma_v^2} - \frac{(s-\eta_s)^2}{2\sigma_s^2}\right]$ with respect to s yields,

$$\hat{\mathbf{s}}_{\mathrm{MAP}} = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_s^2}\eta_s + \frac{\sigma_s^2}{\sigma_v^2 + \sigma_s^2} z$$

When the noise power is much less than the signal power, i.e., $\sigma_v^2 \ll \sigma_s^2$, we have

$$\hat{\mathbf{s}}_{\mathrm{MAP}} = z = \hat{\mathbf{s}}_{\mathrm{ML}}.$$

(It is implied that there is no *a priori* information about s)

Casino example: revisited

Recall the transition matrix is

$$A = \left[\begin{array}{ccc} & F & L \\ F & 0.95 & 0.05 \\ L & 0.1 & 0.9 \end{array}\right]$$

and the emission probability matrix is

$$B = \left[\begin{array}{ccccccc} & 1 & 2 & 3 & 4 & 5 & 6 \\ F & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ L & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{2} \end{array}\right]$$

Casino example: revisited

Denote Y as the number of dies, and X the status of the die, i.e., $X = 0$ means loaded and $X = 1$ indicates fair. If we have the observation $Y = 6$, then we can use maximum likelihood and maximum a posteriori estimate to estimate X.

$$P(Y{=}6|X{=}0){=}\tfrac{1}{2},P(Y{=}6|X{=}1){=}\tfrac{1}{6}$$

Hence $\hat{X}_{\mathrm{ML}} = 0$.

$$P(X{=}0|Y{=}6){=}\tfrac{P(X=0)P(Y=6|X=0)}{P(Y=6)},P(X{=}1|Y{=}6){=}\tfrac{P(X=1)P(Y=6|X=1)}{P(Y=6)},$$

Suppose we have

$$P(X{=}0){=}\tfrac{1}{3},P(X{=}1){=}\tfrac{2}{3}$$

As $\frac{P(X=0)P(Y=6|X=0)}{P(Y=6)} = \frac{1/6}{P(Y=6)} > \frac{P(X=1)P(Y=6|X=1)}{P(Y=6)} = \frac{1/9}{P(Y=6)}$, we have $\hat{X}_{MAP} = 0$, i.e., Cheating happens.

Casino example: revisited

If $Y = (6,2)$, then we have

$$P(Y = (6,2)|X = (0,1)) = 1/12 \quad P(Y = (6,2)|X = (1,1)) = 1/36$$
$$P(Y = (6,2)|X = (0,0)) = 1/20 \quad P(Y = (6,2)|X = (1,0)) = 1/60$$

and $\hat{X}_{\mathrm{ML}} = (0,1)$. On the other hand, we calculate $P(X|Y = (6,2))$, and have

$$P(X = (0,1)|Y = (6,2)) = \frac{1/3 \cdot 0.1 \cdot 1/12}{P(Y=(6,2))} \quad P(X = (1,1)|Y = (6,2)|) = \frac{2/3 \cdot 0.95 \cdot 1/36}{P(Y=(6,2))}$$
$$P(X = (0,0)|Y = (6,2)) = \frac{1/3 \cdot 0.9 \cdot 1/20}{P(Y=(6,2))} \quad P(X = (1,0)|Y = (6,2)) = \frac{2/3 \cdot 0.05 \cdot 1/60}{P(Y=(6,2))}$$

Then $\hat{X}_{\mathrm{MAP}} = (1,1)$.

Is there any cheating?

Contents

- Formulating the problem
- Maximum Likelihood Estimation
- Maximum a posteriori Estimation
- Naive Bayes and logistic regression
- Minimum Mean-Square Error Estimation

Reference

- http://www.cs.cmu.edu/~tom/mlbook-chapter-slides.html
- http://www.cs.cmu.edu/~awm/10701/

An example for classification

Day	Outlook	Temperature	Humidity	Wind	Play
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rainy	Mild	High	Weak	Yes
D5	Rainy	Cool	Normal	Weak	Yes
D6	Rainy	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rainy	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rainy	Mild	High	Strong	No

Bayes Classifier

Training data: $X = [Outlook, Temperature, Humidity, Wind]$,

$Y = PlayTennis$

How to estimating $P(Y|X)$?

- Learning = estimating $P(X|Y), P(Y)$
- Classification = using Bayes rule to calculate $P(Y|X^{new})$
- How shall we represent $P(X|Y), P(Y)$?
- How many parameters must we estimate?

Bayes Classifier

Suppose $X = [X_1, ...X_n]$ where X_i and Y are boolean RVs

- For each instance, we need to estimate $2(2^n - 1)$ such parameters

$$\theta_{ij} = P(X = x_i | Y = y_j)$$

  in which x_i is a n-element vector.

- If X is a vector containing 30 Boolean features, then we will need to estimate more than 3 billion parameters!

Naive Bayes Classifier

Along with decision trees, neural networks, nearest nbr, one of the most practical learning methods.

When to use

- Moderate or large training set available
- Attributes that describe instances are **conditionally independent** given classification

Successful applications:

- Diagnosis
- Classifying text documents

Conditionally independence in Naive Bayes

Suppose $X = [X_1, \ldots, X_n]$ and Y is discrete-valued, conditional independence implies that

$$P(X1 \cdots X_n | Y) = \prod_i P(X_i | Y)$$

i.e., X_i and X_j are conditionally independent given Y, for all $i \neq j$.

Alternatively, X is conditionally independent of Y given Z, if the probability distribution governing X is independent of the value of Y, given the value of Z, i.e.,

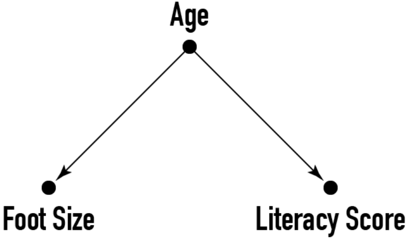
$$P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_k), \forall i, j, k$$

Conditionally independent: example

- $Thunder$ and $Rain$ is not independent, but conditionally independent given $Lightning$.

$$P(Thunder|Rain, Lightning) = P(Thunder|Lightning)$$

- F–Footsize, L–Literacy score, F and L are not independent



- F and L are conditionally independent given age A, i.e.,

$$P(F|L, A) = P(F|A)$$

Naive Bayes Classifier

Naive Bayes uses the assumption that X_i are conditionally independent given Y

then

$$P(X_1, X_2|Y) = P(X_1|X_2, Y)P(X_2|Y) = P(X_1|Y)P(X_2|Y)$$

$$P(X_1, \ldots, X_n|Y) = \prod_i P(X_i|Y)$$

- How many parameters need now for $P(X|Y), P(Y)$?
- We need only $2n$ parameters to define $P(X_k = x_{ik}|Y = y_j)$.

$$P(X = x_i|Y = y_j) = \prod_{k=1}^{n} P(X_k = x_{ik}|Y = y_j)$$

Naive Bayes Classifier

- Bayes rule:

$$P(Y = y_j | X_1, \ldots, X_n) = \frac{P(Y = y_j) P(X_1, \ldots, X_n | Y = y_j)}{\sum_m P(Y = y_m) P(X_1, \ldots, X_n | Y = y_m)}$$

- Assuming conditional independence

$$P(Y = y_j | X_1, \ldots, X_n) = \frac{P(Y = y_j) \prod_i P(X_i | Y = y_j)}{\sum_m P(Y = y_m) \prod_i P(X_i | Y = y_m)}$$

- So, classification rule for $X^{\text{new}} = [x_1^{\text{new}}, \ldots, x_i^{\text{new}}]$ is

$$Y^{\text{new}} \leftarrow \arg\max_{y_j} P(Y = y_j) \prod_i P(X_i = x_i^{\text{new}} | Y = y_j)$$

Naive Bayes: Example

Consider *PlayTennis*, and new instance

$$\langle Outlk = sun, Temp = cool, Humid = high, Wind = strong \rangle$$

Want to compute:

$$Y^{new} = \mathrm{argmax}_j P(Y = y_j) \prod_i P(X_i = x_{ik} | Y = y_j)$$

$$P(y)\ P(sun|y)\ P(cool|y)\ P(high|y)\ P(strong|y) = .005$$

$$P(n)\ P(sun|n)\ P(cool|n)\ P(high|n)\ P(strong|n) = .021$$

$$\rightarrow Y^{new} = n$$

Day	Outlook	Temperature	Humidity	Wind	Play
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rainy	Mild	High	Weak	Yes
D5	Rainy	Cool	Normal	Weak	Yes
D6	Rainy	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rainy	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rainy	Mild	High	Strong	No

Naive Bayes Algorithm

```
Naive_Bayes_Learn($examples$)
    For each target value $y_j$
        estimate $\pi_j = P(Y = y_j)$
        For each attribute value $x_{ik}$ of each attribute $X_i$
            estimate $\theta_{ijk} = P(X_i = x_{ik} | Y = y_j)$
```

Classify_New_Instance(x)

$$Y^{new} = \mathrm{argmax}_{y_j} P(Y = y_j) \prod_i P(X_i | Y = y_j)$$

$$Y^{new} = \mathrm{argmax}_{y_j} \pi_j \prod_i \theta_{ijk}$$

Parameters must sum to 1!

Estimating Parameters: Y, X_i discrete-valued

Maximum likelihood estimates:

$$\pi_j = P(Y = y_j) = \frac{\#D\{Y = y_j\}}{|D|}$$

$$\theta_{ijk} = P(X_i = x_{ik}|Y = y_j) = \frac{\#D\{X_i = x_{ik}, Y = y_j\}}{\#D\{Y = y_j\}}$$

MAP estimates (Laplace smoothing for the case $l = 1$)

$$\pi_j = P(Y = y_j) = \frac{\#D\{Y = y_j\} + l}{|D| + lR}$$

$$\theta_{ijk} = P(X_i = x_{ik}|Y = y_j) = \frac{\#D\{X_i = x_{ik}, Y = y_j\} + l}{\#D\{Y = y_j\} + lM}$$

- $\#D()$ denotes the number of items in data set D.
- MAP estimate for θ_{ijk} if we assume a Dirichlet prior distribution over the θ_{ijk} parameters, with equal-valued parameters

Naive Bayes: Subtleties

1. Conditional independence assumption is often violated

$$P(X_1, X_2 \ldots X_n | Y_j) = \prod_i P(X_i | Y_j)$$

- ...but it works surprisingly well anyway. Note don't need estimated posteriors $P(Y_k|X)$ to be correct; need only that

$$\mathrm{argmax}_{y_k} P(Y = y_k) \prod_i P(X_i | Y = y_k) =$$

$$\mathrm{argmax}_{y_k} P(Y = y_k) P(X_1 \ldots, X_n | Y = y_k)$$

- see [Domingos & Pazzani, 1996] for analysis
- Naive Bayes posteriors often unrealistically close to 1 or 0

Naive Bayes: Subtleties

2. what if none of the training instances with target value y_k have attribute value x_{ij}? Then

$$P(X_i = x_{ij}|Y = y_k) = 0, \text{ and...} P(Y = y_k)\prod_i P(X_i = x_{ij}|Y = y_k) = 0$$

MAP Estimate mentioned before!

Naive Bayes: Subtleties

3. What if we have continuous X_i?

   For example, in image classification, X_i is the i-th pixel, Gaussian Naive Bayes (GNB) assumes:

$$p(X_i = x|Y = y_k) = \frac{1}{\sqrt{2\pi}\sigma_{ik}} e^{-\frac{(x-\mu_{ik})^2}{2\sigma_{ik}^2}}$$

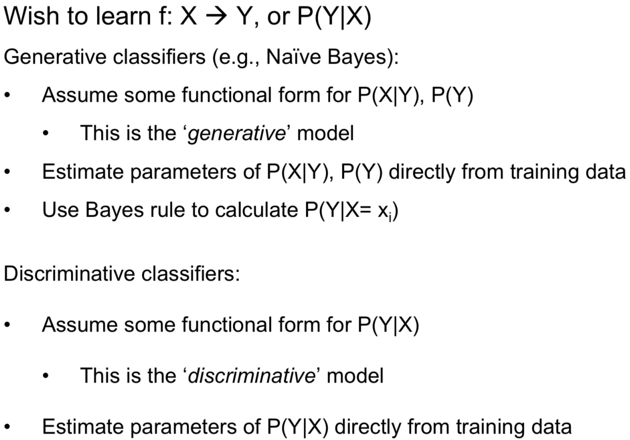
   Sometimes assume variance

   - is independent of Y (i.e., σ_i)
   - or independent of X_i (i.e., σ_k)
   - or both (i.e., σ)

Application of Naive Bayes

- Credit scoring
- Medical data classification
- Classify which emails are spam
- Classify which emails are meeting invites
- Classify which web pages are student home pages (Recommendation system)
- Sentiment analysis

Generative vs. Discriminative Classifiers



Wish to learn f: X → Y, or P(Y|X)

Generative classifiers (e.g., Naïve Bayes):

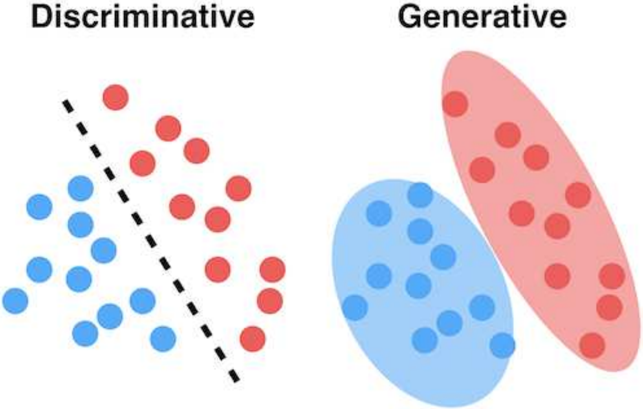
- Assume some functional form for P(X|Y), P(Y)
  - This is the '*generative*' model
- Estimate parameters of P(X|Y), P(Y) directly from training data
- Use Bayes rule to calculate P(Y|X= x_i)

Discriminative classifiers:

- Assume some functional form for P(Y|X)
  - This is the '*discriminative*' model
- Estimate parameters of P(Y|X) directly from training data

In General, A Discriminative model models the decision boundary between the classes. A Generative Model explicitly models the actual distribution of each class.

Generative vs. Discriminative Classifiers



Examples of generative and discriminative classifiers

Generative classifiers

- Naive Bayes
- Bayesian networks
- Markov random fields
- Hidden Markov Models (HMM)

Discriminative Classifiers

- Logistic regression
- Support Vector Machine
- Traditional neural networks
- Nearest neighbour
- Conditional Random Fields (CRF)s

From Gaussian Naive Bayes to Logistic Regression

- Consider learning $f : X \to Y$, where
  - X is vector of real-valued features, $X = [X_1, \ldots, X_n]^T$
  - Y is Boolean
- We could use a Gaussian Naive Bayes classifier
  - assume all X_i are conditionally independent given Y
  - model the probability density $p(X_i|Y = y_k)$ as Gaussian $N(\mu_{ik}, \sigma_i)$
- What does that imply about the form of $P(Y|X)$?

Mixed joint density

The "mixed joint density" may be defined where one or more random variables are continuous and the other random variables are discrete. With one variable of each type

$$f_{X,Y}(x,y) = f_{X|Y}(x \mid y)\mathrm{P}(Y=y) = \mathrm{P}(Y=y \mid X=x) f_X(x).$$

Formally, $f_{X,Y}(x,y)$ is the probability density function of (X,Y) with respect to the product measure on the respective supports of X and Y, and we have the joint cumulative distribution function

$$F_{X,Y}(x,y) = \sum_{t \le y} \int_{-\infty}^{x} f_{X,Y}(s,t) ds.$$

Derive form for $P(Y|X)$ for continuous X_i

$$P(Y=1|X) = \frac{P(Y=1)f(X|Y=1)}{f(x)}$$

$$P(Y=0|X) = \frac{P(Y=0)f(X|Y=0)}{f(x)}$$

$$\frac{P(Y=1|X)}{P(Y=0|X)} = \frac{P(Y=1)f(X|Y=1)}{P(Y=0)f(X|Y=0)}$$

From Gaussian Naive Bayes to Logistic Regression

Assume $P(Y=1)=\pi$, $P(Y=0)=1-\pi$, and

$P(X_i|Y=j)=\frac{1}{\sqrt{2\pi}\sigma_i}e^{-\frac{(x_i-\mu_{ij})^2}{2\sigma_i^2}}$, then we have

$$\frac{P(Y=0|X)}{P(Y=1|X)}=\exp(w_0+\sum_{i=1}^{n}w_ix_i)$$

with

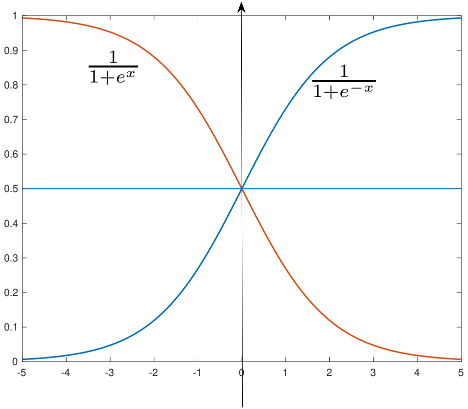
$$w_0=\ln\frac{1-\pi}{\pi}+\sum_i\frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}, w_i=\frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2}$$

which then implies,

$$\ln\frac{P(Y=0|X)}{P(Y=1|X)}=w_0+\sum_{i=1}^{n}w_ix_i$$

Linear classification rule!!

Logistic function



$$Y = 0 \leftarrow x > 0 (w_0 + \sum_i w_i x_i > 0)$$

Estimating parameters for Logistic regression

- The value of the weights w_i can be provided by the parameters estimated by the GNB classifier
- The form of $p(Y|X)$ assumed by Logistic regression holds in many problem settings beyond the GNB problem
- In many cases we may suspect the GNB assumptions are not perfectly satisfied
- Estimate the w_i parameters directly from the data!

Estimating parameters for logistic regression

- Choose parameters $W = [w_0, \ldots, w_n]$ to maximize conditional likelihood of training data

$$P(Y=1|X) = \frac{\exp(w_0 + \sum_{i=1}^n w_i x_i)}{1 + \exp(w_0 + \sum_{i=1}^n w_i x_i)}$$

$$P(Y=0|X) = \frac{1}{1 + \exp(w_0 + \sum_{i=1}^n w_i x_i)}$$

- Given training data $D = \{(X^1, Y^1), \ldots, (X^L, Y^L)\}$
- Data conditional likelihood $= \prod_l P(Y^l|X^l, W)$

$$W \leftarrow \arg\max_W \ln \prod_l P(Y^l|X^l, W)$$

Expressing Conditional Log Likelihood

$$l(W) = \ln \prod_l P(Y^l|X^l, W) = \sum_l \ln P(Y^l|X^l, W)$$

Flip the assignment of the boolean variable Y

$$P(Y = 0|X, W) = \frac{1}{1 + \exp(w_0 + \sum_i w_i X_i)},$$

$$P(Y = 1|X, W) = \frac{\exp(w_0 + \sum_i w_i X_i)}{1 + \exp(w_0 + \sum_i w_i X_i)}$$

$$\begin{aligned} l(W) &= \sum_l \left[Y^l \ln P(Y^l = 1|X^l, W) + (1 - Y^l) \ln P(Y^l = 0|X^l, W)\right] \\ &= \sum_l \left[Y^l \ln \tfrac{P(Y^l=1|X^l,W)}{P(Y^l=0|X^l,W)} + \ln P(Y^l = 0|X^l, W)\right] \\ &= \sum_l \left[Y^l(w_0 + \sum_{i=1}^n w_i X_i^l) - \ln(1 + \exp(w_0 + \sum_{i=1}^n w_i X_i^l))\right] \end{aligned}$$

Maximizing Conditional Log Likelihood

$$l(W) = \sum_{l} \left[Y^l (w_0 + \sum_{i=1}^{n} w_i X_i^l) - \ln(1 + \exp(w_0 + \sum_{i=1}^{n} w_i X_i^l)) \right]$$

- Good news: $l(W)$ is a concave function of W
- Bad news: no closed-form solution to maximize W

Gradient-based method! (Gradient-ascent)

Regularization in Logistic Regression

How about MAP?

- On common approach is to define priors on W
  - Normal distribution, zero mean, identity covariance
- Helps avoid very large weights and overfitting
- MAP estimate

$$W \leftarrow \mathrm{argmax}_W \ln P(W|\{X^l, Y^l\})$$

$$W \leftarrow \mathrm{argmax}_W \ln P(W)P(Y^l|X^l, W)$$

Regularization in Logistic Regression

$$\sum_l \ln P(Y^l|X^l, W) + \ln P(W)$$

if $P(W)$ is a zero mean Gaussian distribution, then $\ln P(W)$ yields a term proportional to $\|W\|^2$.

$$W \leftarrow \mathrm{argmax}_W \sum_l \ln P(Y^l|X^l, W) - \frac{\lambda}{2}\|W\|^2$$

Logistic Regression for functions with many discrete values

- Logistic regression in more general case, where $Y \in \{Y_1, \ldots, Y_R\}$: learn $R-1$ sets of weights

  for $k < R$

$$P(Y = y_k|X) = \frac{\exp(w_{k0} + \sum_{i=1}^{n} w_{ki} X_i)}{1 + \sum_{j=1}^{R-1} \exp(w_{j0} + \sum_{i=1}^{n} w_{ji} X_i)}$$

  for $k = R$

$$P(Y = y_R|X) = \frac{1}{1 + \sum_{j=1}^{R-1} \exp(w_{j0} + \sum_{i=1}^{n} w_{ji} X_i)}$$

Softmax layer in CNN!

Relationship between Gaussian Naive Bayes (GNB) classifiers and Logistic Regression

- Generative and Discriminative classifiers
- When the GNB modeling assumptions do not hold, Logistic Regression and GNB typically learn different classifier function
  - the asymptotic (#samples $\to \infty$) classification accuracy for logistic regression is often better than that of GNB
  - the GNB assumption do not need to be satisfied for Logistic Regression
- GNB and Logistic Regression converges toward their asymptotic accuracies at different rates
  - when GNB parameter estimates converge in order $\log n$ examples
  - Logistic Regression requiring order n examples
  - In general, when many training examples are available, we choose Logistic Regression, otherwise choose GNB.

Contents

- Formulating the problem
- Maximum Likelihood Estimation
- Maximum a posteriori Estimation
- Naive Bayes and logistic regression
- Minimum Mean-Square Error Estimation

Minimum Mean-square error

$$\mathrm{MSE} = \mathrm{E}[\mathrm{E}[\tilde{\mathbf{s}}^2|\mathbf{z}]] = \mathrm{E}[\mathrm{E}[(\mathbf{s} - \hat{\mathbf{s}})^2|\mathbf{z}]] = \mathrm{E}[(\mathbf{s} - \hat{\mathbf{s}})^2]$$

The MSE gives the average power of the error.

Given the RV $\mathbf{z}$, the MMSE estimate $\hat{s}_{\mathrm{MMSE}}$ of $\hat{s}$ is the conditional expectation

$$\hat{s}_{\mathrm{MMSE}} = E[\mathbf{s}|\mathbf{z}].$$

Properties of MMSE estimate

- MMSE estimate $\hat{\mathbf{s}}_{\text{MMSE}}$ is unique;
- MMSE estimate requires information about $\mathbf{s}$, is another type of Baysian estimation;
- MMSE estimate $\hat{\mathbf{s}}_{\text{MMSE}}$ is unbiased, i.e.,

$$E(\hat{\mathbf{s}}) = E(\mathbf{s}) \text{ or } E(\tilde{\mathbf{s}}) = 0$$

- Generalization to a finite number of measurements $\mathbf{z}(1), \ldots, \mathbf{z}(n)$,

$$\hat{\mathbf{s}}_{\text{MMSE}} = E[\mathbf{s}|\mathbf{z}(1), \ldots, \mathbf{z}(n)], \quad E[\mathbf{s} - \hat{\mathbf{s}}_{\text{MMSE}}] = 0$$

MMSE estimate with Gaussian noise

Again consider the additive-noise case

$$\mathbf{z} = \mathbf{s} + \mathbf{v}$$

with $\mathbf{v} \sim \mathcal{N}(0, \sigma_v^2)$ and $\mathbf{s} \sim \mathcal{N}(\bar{s}, \sigma_s^2)$. Assume that $\mathbf{s}$ and $\mathbf{v}$ are independent.

$$E[\mathbf{z}] = E[\mathbf{s}] = \bar{s},$$

$$\mathrm{Var}[\mathbf{z}] = \mathrm{Var}[\mathbf{s}] + \mathrm{Var}[\mathbf{v}] = \sigma_s^2 + \sigma_v^2$$

The pdf of $\mathbf{z}$ is

$$f_{\mathbf{z}}(z) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma_s^2 + \sigma_v^2}} \exp\left[-\frac{(z-\bar{s})^2}{2(\sigma_s^2 + \sigma_v^2)}\right]$$

$$f_{\mathbf{s}}(s|\mathbf{z}=z)=\frac{1}{2\pi\sigma_s\sigma_v f_{\mathbf{z}}(z)}\exp\left\{-\left[\frac{(z-s)^2}{2\sigma_v^2}+\frac{(s-\bar{s})^2}{2\sigma_s^2}\right]\right\}$$

$$f_{\mathbf{s}}(s|\mathbf{z}=z)=\frac{1}{\sqrt{2\pi}\sqrt{\frac{\sigma_s^2\sigma_v^2}{\sigma_s^2+\sigma_v^2}}}\exp\left[-\frac{(s-\hat{s}_{\mathrm{MAP}})^2}{2\frac{\sigma_s^2\sigma_v^2}{\sigma_s^2+\sigma_v^2}}\right]$$

$$\hat{s}_{\mathrm{MMSE}}=E[\mathbf{s}|\mathbf{z}=z]=\hat{s}_{\mathrm{MAP}}=\bar{s}+\frac{\sigma_s^2}{\sigma_v^2+\sigma_s^2}(z-\bar{s})$$

If $\mathbf{s}$ and $\mathbf{v}$ are uncorrelated, then the MMSE estimate of $\mathbf{s}$ is identical to the MAP estimate.

The Orthogonality Principle

Orthogonality Principle

The error $\mathbf{s} - E[\mathbf{s}|\mathbf{z}]$ is orthogonal to every function $\gamma(\mathbf{z})$, i.e.,

$$E[(\mathbf{s} - E[\mathbf{s}|\mathbf{z}])\gamma(\mathbf{z})] = 0.$$

Sketch of proof. $E[(\mathbf{s} - E[\mathbf{s}|\mathbf{z}])\gamma(\mathbf{z})] = E\left\{E[(\mathbf{s} - E[\mathbf{s}|\mathbf{z}])\gamma(\mathbf{z})|\mathbf{z}]\right\} = E\left\{E[(\mathbf{s} - E[\mathbf{s}|\mathbf{z}])|\mathbf{z}]\gamma(\mathbf{z})\right\}$

Necessary and Sufficient condition for an MMSE estimate

The estimate given by $\hat{\mathbf{s}} = \alpha(\mathbf{z})$ is the MMSE estimate of $\mathbf{s}$ given $\mathbf{z}$ if and only if the error $\mathbf{s} - \alpha(\mathbf{z})$ is orthogonal to every function $\gamma(\mathbf{z})$; that is

$$E[(\mathbf{s} - \alpha(\mathbf{z}))\gamma(\mathbf{z})] = 0.$$

$\hat{\mathbf{s}} = \alpha(\mathbf{z})$ is the MMSE estimate $\Longleftrightarrow (\mathbf{s} - \alpha(\mathbf{z})) \perp \gamma(\mathbf{z})$

Linear MMSE

the MMSE estimate is a conditional expectation.

$f_{\mathbf{s}}(s|\mathbf{z} = z)$ is difficult to be find $\Longleftrightarrow$ both MAP and MMSE estimation are difficult.

Solution: restrict the estimation problem to produce a tractable solution for α, i.e., trading overall optimality for tractability.

$$\hat{\mathbf{s}} = \lambda \mathbf{z}$$

Linear MMSE

Linear MMSE (LMMSE) estimation problem:

$$\min_{\lambda}\ \mathrm{MSE} = E[(\mathbf{s} - \lambda\mathbf{z})^2] = E[\mathbf{s}^2 - 2\lambda\mathbf{s}\mathbf{z} + \lambda^2\mathbf{z}^2]$$

Taking the partial derivative with respect to λ, setting the result equal to zero gives

$$-2E(\mathbf{s}\mathbf{z}) + 2\lambda E(\mathbf{z}^2) = 0, \lambda = \frac{E(\mathbf{s}\mathbf{z})}{E(\mathbf{z}^2)}$$

The LMMSE estimate is given by

$$\hat{s}_{\mathrm{LMMSE}} = \alpha(\mathbf{z}) = \frac{E(\mathbf{s}\mathbf{z})}{E(\mathbf{z}^2)} \cdot \mathbf{z}$$

Advantages over the former estimates

- do not require knowledge about any likelihood function or densities;
- only need the second-order moments $E[\mathbf{sz}]$ and $E[\mathbf{z}^2]$;
- can estimate $E[\mathbf{sz}]$ and $E[\mathbf{z}^2]$ from experimental training data $(s_i, z_i)_{i=1}^N$, i.e.,

$$E(\mathbf{sz}) \approx \frac{1}{N} \sum_{i=1}^{N} s_i z_i$$

and

$$E(\mathbf{z}^2) \approx \frac{1}{N} \sum_{i=1}^{N} z_i^2$$

Example

Consider again the Gaussian additive noise problem.

$$\mathbf{z} = \mathbf{s} + \mathbf{v}$$

with $\mathbf{v} \sim \mathcal{N}(0, \sigma_v^2)$ and $\mathbf{s} \sim \mathcal{N}(\bar{s}, \sigma_s^2)$. Assume that $\mathbf{s}$ and $\mathbf{v}$ are uncorrelated.

$$\hat{\mathbf{s}}_{\mathrm{LMMSE}} = \frac{E(\mathbf{s}\mathbf{z})}{E(\mathbf{z}^2)}\mathbf{z} = \frac{\bar{s}^2 + \sigma_s^2}{\bar{s}^2 + \sigma_s^2 + \sigma_v^2} \cdot \mathbf{z}$$

which is different from $\hat{\mathbf{s}}_{\mathrm{MMSE}}$ and is biased.

Orthogonality principle for LMMSE estimation

Orthogonality principle for LMMSE estimation

Let $\alpha(\mathbf{z})$ be the LMMSE estimate of $\mathbf{s}$ given $\mathbf{z}$. Then the error $\mathbf{s} - \alpha(\mathbf{z})$ is orthogonal to every linear function $\gamma(\mathbf{z})$, i.e.,

$$E[(\mathbf{s} - \alpha(\mathbf{z}))\gamma(\mathbf{z})] = 0.$$

Sketch of proof. Assume $\gamma(\mathbf{z}) = \beta\mathbf{z}$

Orthogonality principle for vector RVs

Let $\mathbf{s} \in \mathbb{R}^m$ and $\mathbf{z} \in \mathbb{R}^q$. Assume the LMMSE takes the form $\hat{\mathbf{s}} = M\mathbf{z}$, where M is an $m \times q$ matrix to be determined.

Assume $P = E[(\mathbf{s} - \hat{\mathbf{s}})(\mathbf{s} - \hat{\mathbf{s}})^T]$, then

$$\mathrm{MSE} = \mathrm{tr}(P) = E[(\mathbf{s} - \hat{\mathbf{s}})^T(\mathbf{s} - \hat{\mathbf{s}})]$$

Matrix derivative

$f(A) : \mathbb{R}^{m\times n} \to 1$, the Jacobian matrix ($A = [A_1, A_2, \ldots, A_n]$),

$$\nabla_A f = \frac{\partial f}{\partial A} = \begin{bmatrix} \frac{\partial f}{\partial a_{11}} & \frac{\partial f}{\partial a_{12}} & \cdots & \frac{\partial f}{\partial a_{1n}} \\ \frac{\partial f}{\partial a_{21}} & \frac{\partial f}{\partial a_{22}} & \cdots & \frac{\partial f}{\partial a_{2n}} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f}{\partial a_{n1}} & \frac{\partial f}{\partial a_{n2}} & \cdots & \frac{\partial f}{\partial a_{nn}} \end{bmatrix} = \left[\frac{\partial f}{\partial A_1}, \frac{\partial f}{\partial A_2}, \ldots, \frac{\partial f}{\partial A_n}\right]$$

A list of derivatives

$$\frac{\partial(Ax)}{\partial x} = A, \frac{\partial(a^TAb)}{\partial A} = ab^T, \frac{\partial(a^TA^Tb)}{\partial A} = ba^T$$

$$\frac{\partial \mathrm{tr}(C^TAB^T)}{\partial A} = \frac{\partial \mathrm{tr}(BA^TC)}{\partial A} = CB$$

$$\frac{\partial^2}{\partial x \partial x^T}(Ax+b)^TC(Dx+e) = A^TCD + D^TC^TA$$

$$\frac{\partial^2}{\partial x \partial x^T}(x^TCx) = C + C^T$$

$$\frac{\partial}{\partial x}(Ax+b)^TC(Dx+e) = A^TC(Dx+e) + D^TC^T(Ax+b)$$

$$\frac{\partial}{\partial A}(Aa+b)^TC(Aa+b) = (C+C)^T(Aa+b)a^T$$

Solution of the LMMSE

Differentiating the MSE with respect to M yields,

$$\frac{\partial \mathrm{tr}(P)}{\partial M} = -2E[\mathbf{s}\mathbf{z}^T] + 2ME[\mathbf{z}\mathbf{z}^T]$$

Hence

$$M = E(\mathbf{s}\mathbf{z}^T)[E(\mathbf{z}\mathbf{z}^T)]^{-1}$$

Thus the LMMSE estimate of $\hat{s}$ given $\mathbf{z}$ is

$$\hat{\mathbf{s}}_{\mathrm{LMMSE}} = E(\mathbf{s}\mathbf{z}^T)\left[E(\mathbf{z}\mathbf{z}^T)\right]^{-1}\mathbf{z}$$

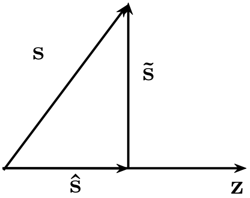
Orthogonality principle for vector RVs

Let $\mathbf{s} \in \mathbb{R}^m$ and $\mathbf{z} \in \mathbb{R}^q$ be jointly distributed random vectors, let $\hat{\mathbf{s}} = \alpha(\mathbf{z})$ be the LMMSE estimate of $\mathbf{s}$ given $\mathbf{z}$. Then the estimation error $\mathbf{s} - \hat{\mathbf{s}}$ is orthogonal to $\mathbf{z}$, i.e.,

$$E[(\mathbf{s} - \hat{\mathbf{s}})\mathbf{z}^T] = \mathbf{0}.$$

Sketch of proof. Just use the expression for $\hat{\mathbf{s}}$.

Illustration of orthogonality principle



Necessary and sufficient condition for orthogonality

Let $\alpha(\mathbf{z})$ be a linear estimator of $\mathbf{s}$ given $\mathbf{z}$. Then α minimizes the MSE if and only if the error $\mathbf{s} - \alpha(\mathbf{z})$ is orthogonal to the measurement $\mathbf{z}$,

$$E\left\{[\mathbf{s} - \alpha(\mathbf{z})]\mathbf{z}^T\right\} = 0$$

We can use this to find the optimum linear estimator.

Overall optimality

- When $\mathbf{s}$ and $\mathbf{z}$ are zero-mean, jointly Gaussian, the LMMSE estimate is also the optimal MMSE estimate.
  - Suppose that $\mathbf{s}$ and $\mathbf{z}$ have a zero-mean bivariate Gaussian distribution with covariance matrix P given by

$$P = \begin{bmatrix} \sigma_s^2 & \mathrm{Cov}(\mathbf{s}, \mathbf{z}) \\ \mathrm{Cov}(\mathbf{z}, \mathbf{s}) & \sigma_z^2 \end{bmatrix}$$

  - the conditional density function $f_{\mathbf{s}}(s|\mathbf{z} = z)$ is given by

$$f_{\mathbf{s}}(s|\mathbf{z} = z) = \frac{1}{\sqrt{2\pi}\sigma_s\sqrt{(1-\rho^2)}} \exp\left\{ -\frac{1}{2\sigma_s^2(1-\rho^2)} \left(s - \frac{E(\mathbf{sz})}{E(\mathbf{z}^2)} z \right)^2 \right\}$$

  - $\hat{s}_{\mathrm{MMSE}} = \hat{s}_{\mathrm{LMMSE}} = \frac{E(\mathbf{sz})}{E(\mathbf{z}^2)}\mathbf{z}$

Comparison of different estimators

	Maximum likelihood (ML)	Maximum *a posteriori* (MAP)
Motivation	Given z, what value of $\mathbf{s}$ is most likely to have produced z?	Given z, what value of $\mathbf{s}$ is most likely to have occured?
Objective	Maximize the likelihood function $f_{\mathbf{z}}(z\|\mathbf{s}=s)$	maximize the conditional density $f_{\mathbf{s}}(s\|\mathbf{z}=z)$ via Bayes rule, equivalently maximize $f_{\mathbf{z}}(z\|\mathbf{s}=s)f_{\mathbf{s}}(s)$.
Esitmate	$\hat{s}_{\mathrm{ML}} = \mathrm{argmax} f_{\mathbf{z}}(z\|\mathbf{s}=s)$	$\hat{s}_{\mathrm{MAP}} = \mathrm{argmax} f_{\mathbf{z}}(z\|\mathbf{s}=s)f_{\mathbf{s}}(s)$
Required knowledge	likelihood function $f_{\mathbf{z}}(z\|\mathbf{s}=s)$	Density function $f_{\mathbf{s}}(s\|z)$ (or $f_{\mathbf{z}}(z\|\mathbf{s}=s)$ and $f_{\mathbf{s}}(s)$)

Comparison of different estimators

	Minimum mean-square error (MMSE)	Linear MMSE (LMMSE)
Motivation	Given z, what estimate of s gives the smallest MSE?	Given z, what linear function $\hat{s} = \lambda\mathbf{z}$ gives the smallest MSE?
Objective	Minimize the MSE $E[(\mathbf{s}-\hat{\mathbf{s}})^2]$	find λ to minimize $E[(\mathbf{s}-\lambda\mathbf{z})^2]$.
Esitmate	$\hat{\mathbf{s}}_{\mathrm{MMSE}} = E(\mathbf{s}\mid\mathbf{z}) = \int_{-\infty}^{\infty} s f_{\mathbf{s}}(s\mid\mathbf{z})ds$	$\hat{s}_{\mathrm{LMMSE}} = \lambda\mathbf{z}$, where $\lambda = E[\mathbf{sz}]/E[\mathbf{z}^2]$
Required knowledge	Density $f_{\mathbf{s}}(s\mid z)$	Cross-correlation of $\mathbf{s}$ and $\mathbf{z}$ $E[\mathbf{sz}]$; second moment of $\mathbf{z}$, $E(\mathbf{z}^2)$