Lecture 25 EE 559 Apr 22, 2024

Machine Learning I: Supervised Methods

B. Keith Jenkins

Announcements

- Homework 8 is due today
- Project Reports and code are due Friday 4/26
  - See piazza for late submission policy
- Final exam is Wed., May 1

Today's lecture

- Summary and orientation for estimating $p(x \mid S_k)$ for classification
- Parameter estimation for classification
  - Problem set-up and notation
  - Ad hoc estimates
  - View 1: parameters are deterministic
    - Maximum likelihood estimate (MLE)
  - View 2: parameters are random
    - MAP estimate
    - Prior on parameters
    - Use in a classifier (2 approaches)
      - MAP as point estimate
      - Full Bayesian (integrating)
- d.o.f. and constraints in parameter estimation

Summary so far (probabilistic classification) (2-class case)

L22: Bayes dec. rule for 2-class:

$$p(\underline{x}|S_1)P(S_1) \underset{\Gamma_2}{\overset{\Gamma_1}{\gtrless}} p(\underline{x}|S_2)P(S_2)$$

Estimate priors $\hat{P}(S_i)$

Ex: frequency of occurrence $\hat{P}(S_i) = \frac{N_i}{N}$ → $\frac{N_1}{N}$, $\frac{N_2}{N}$

L23: Estimate $p(\underline{x}|S_i)$

1. Density estimation
   - Ex: • kNN
     • KDE

Now: **2. Parameter estimation**

- Ex: • maximum likelihood est. (MLE)
  - max. a posteriori (MAP) (as a point estimate)
  - Full Bayesian est. (integrating over parameter posterior)

→ $\hat{p}(\underline{x}|S_1)$, $\hat{p}(\underline{x}|S_2)$

→ $f_1(\underline{x}, \hat{\underline{\theta}}^{(1)}_{MLE})$, $f_2(\underline{x}, \hat{\underline{\theta}}^{(2)}_{MLE})$

→ $f_1(\underline{x}, \hat{\underline{\theta}}^{(1)}_{MAP})$, $f_2(\underline{x}, \hat{\underline{\theta}}^{(2)}_{MAP})$

→ $p(\underline{x}|S_1, \underline{\mathcal{Z}}_1)$, $p(\underline{x}|S_2, \underline{\mathcal{Z}}_2)$

↑ data labelled S_1 ↑ data labelled S_2

$f_k(\underline{x}, \underline{\theta}^{(k)})$ = known or assumed functional form of $p(\underline{x}|S_k)$.

Parameter Estimation

Motivation:

1. Reduce d.o.f. by making more assumptions
2. Include any knowledge of $p(\underline{x}|S_k)$

(1) Assume: $p(\underline{x}|S_k) = f_k(\underline{x}, \underline{\theta}_k)$ ← functional form of f_k is known or assumed.

$\underline{\theta}_k$ are parameters that are unknown.

Ex: $p(\underline{x}|S_k) = \mathcal{N}(\underline{x}|\underline{m}_k, \underline{\underline{\Sigma}}_k) \triangleq f_k(\underline{x}, \underline{\theta}_k)$

with $\underline{m}_k$ unknown, and/or $\underline{\underline{\Sigma}}_k$ unknown.

→ Unknown parameters $\underline{\theta}_k$ are learned from the training data.

Problem set-up and notation

Let $\hat{\underline{\theta}}_k$ be an estimate of $\underline{\theta}_k$. (Parameters of $p(\underline{x}|S_k)$)

Let $\underline{x}_1^{(k)}, \underline{x}_2^{(k)}, \cdots, \underline{x}_n^{(k)}$ be the data pts., drawn i.i.d. from $p(\underline{x}|S_k)$.

Let $\underline{\underline{X}} = [\underline{x}_1\ \underline{x}_2\ \underline{x}_3 \cdots \underline{x}_n]^T$ denote all training data

$\underline{\underline{X}}_k = [\underline{x}_1^{(k)}\ \underline{x}_2^{(k)} \cdots \underline{x}_{n_k}^{(k)}]^T$ denote training data labeled S_k.

Class independence assumption

Assume: to estimate $\underline{\theta}_k$, we only need data $\underline{z}_k$ from S_k.

When considering estimates for only 1 class S_k, will simplify notation:

$$p(\underline{x} \mid \underline{\theta}_k, S_k) \rightarrow p(\underline{x} \mid \underline{\theta})$$

$$\underline{x}_m^{(k)} \rightarrow \underline{x}_m, \quad \underline{z}^{(k)} \rightarrow \underline{z}$$

Is $\underline{\theta}$ deterministic (but unknown), or random?

→ Both views are valid.

View 1: $\underline{\theta}$ is deterministic

$$\hat{\underline{\theta}} = G(\underline{x}_1, \underline{x}_2, \cdots, \underline{x}_n) = G(\underline{z}) = \hat{\underline{\theta}}(\underline{z})$$

Two properties of an estimate $\hat{\underline{\theta}}$

Unbiased

If $E_{\underline{z}}\{\hat{\underline{\theta}}\} \triangleq \int \hat{\underline{\theta}}(\underline{z})\, p(\underline{z})\, d\underline{z} = \underline{\theta}$

Then $\hat{\underline{\theta}}$ is an unbiased estimate of $\underline{\theta}$.

Otherwise, $\hat{\underline{\theta}}$ is biased.

Consistent

If $\lim_{n\to\infty} P\{\|\hat{\underline{\theta}} - \underline{\theta}\| < \epsilon\} = 1$ for every $\epsilon > 0$,

then $\hat{\underline{\theta}}$ is consistent.

("convergence in probability")

Ad hoc estimates

Sample mean

$$\hat{\underline{m}} \triangleq \frac{1}{n} \sum_{i=1}^{n} \underline{x}_i$$

- unbiased?

$$E\{\hat{\underline{m}}\} = E\left\{\frac{1}{n}\Sigma \underline{x}_i\right\} = \frac{1}{n}\Sigma E\{\underline{x}_i\}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \underline{m} = \frac{1}{n}(n\underline{m}) = \underline{m}$$

$\Rightarrow \hat{\underline{m}}$ is unbiased.

- consistent?

Can show $\lim_{n\to\infty} P\{\|\hat{\underline{m}} - \underline{m}\| < \epsilon\} = 1$

$\Rightarrow \hat{\underline{m}}$ is consistent.

Sample covariance

$$\hat{\underline{\underline{\Sigma}}} = a_n \sum_{i=1}^{n} (\underline{x}_i - \hat{\underline{m}})(\underline{x}_i - \hat{\underline{m}})^T$$

$a_n = \frac{1}{n}$ (biased)

$a_n = \frac{1}{n-1}$ (unbiased)

Maximum Likelihood Estimate (MLE)

[Bishop pp. 26 – middle of p. 28]

Let $p(\underline{\mathcal{Z}} \mid \underline{\theta})$ be the likelihood of parameter values $\underline{\theta}$ based on training data $\underline{\mathcal{Z}}$.

The MLE of $\underline{\theta}$ is:

(2) $$\underline{\hat{\theta}}_{MLE} \triangleq \underset{\underline{\theta}}{\operatorname{argmax}} \left\{ p(\underline{\mathcal{Z}} \mid \underline{\theta}) \right\} = \underset{\underline{\theta}}{\operatorname{argmax}} \left\{ \ln p(\underline{\mathcal{Z}} \mid \underline{\theta}) \right\}$$

$\Rightarrow$ $\underline{\hat{\theta}}_{MLE}$ yields the $\underline{\theta}$ that maximizes the probability of obtaining the data that was actually observed.

(3)
$$\therefore \text{ Criterion fcn. } \tilde{J}_{MLE}(\underline{\theta}) = -p(\underline{\mathcal{Z}} \mid \underline{\theta})$$

$$\text{or } J_{MLE}(\underline{\theta}) = -\ln p(\underline{\mathcal{Z}} \mid \underline{\theta}) \quad \text{("negative log likelihood")}$$

$$= -\ln \prod_{i=1}^{n} p(\underline{x}_i \mid \underline{\theta})$$

$$= -\sum_{i=1}^{n} \ln p(\underline{x}_i \mid \underline{\theta}).$$

Is J_{MLE} convex?

$\rightarrow$ Depends on $f_k(\underline{x}, \underline{\theta}_k)$.

Ex:

Let $p(\underline{x}|\underline{\theta}) = \mathcal{N}(\underline{x}|\underline{\mu}, \underline{\underline{\Sigma}})$, $\underline{\mu}$ = unknown, $\underline{\underline{\Sigma}}$ = known.

$$\hat{\underline{\theta}}_{MLE} = \arg\max_{\underline{\theta}} \{\ln p(\underline{\mathcal{X}}|\underline{\theta})\} = \arg\max_{\underline{\mu}} \{\ln p(\underline{x}_1, \cdots, \underline{x}_n|\underline{\mu})\}$$

$$\hat{\underline{\mu}}_{MLE} = \arg\max_{\underline{\mu}} \left\{ \sum_{i=1}^{n} \ln p(\underline{x}_i|\underline{\mu}) \right\}$$

$$= \arg\max_{\underline{\mu}} \left\{ \sum_{i=1}^{n} -\frac{1}{2}\ln(2\pi)^D - \frac{1}{2}\ln|\underline{\underline{\Sigma}}| - \frac{1}{2}(\underline{x}_i - \underline{\mu})^T \underline{\underline{\Sigma}}^{-1}(\underline{x}_i - \underline{\mu}) \right\}$$

$$-\frac{1}{2}\underline{x}_i^T \underline{\underline{\Sigma}}^{-1}\underline{x}_i + \underline{x}_i^T \underline{\underline{\Sigma}}^{-1}\underline{\mu} - \frac{1}{2}\underline{\mu}^T \underline{\underline{\Sigma}}^{-1}\underline{\mu}$$

drop (constants of $\underline{\mu}$): $\uparrow$ $\uparrow$ and $-\frac{1}{2}\underline{x}_i^T \underline{\underline{\Sigma}}^{-1}\underline{x}_i$.

$$\hat{\underline{\mu}}_{ML} = \arg\min_{\underline{\mu}} J, \quad J = -\sum_{i=1}^{n} \left[\underline{\mu}^T \underline{\underline{\Sigma}}^{-1}\underline{x}_i - \frac{1}{2}\underline{\mu}^T \underline{\underline{\Sigma}}^{-1}\underline{\mu} \right]$$

Is J convex? yes ✓ ✓

min J: $\nabla_{\underline{\mu}} J(\underline{\mu}) = \underline{0}$

$$\nabla_{\underline{\mu}}\{\cdot\} = \sum_{i=1}^{n} \left[\underline{\underline{\Sigma}}^{-1}\underline{x}_i - \underline{\underline{\Sigma}}^{-1}\underline{\mu} \right]$$

$$\sum_{i=1}^{n} \left[\underline{\underline{\Sigma}}^{-1}\underline{x}_i - \underline{\underline{\Sigma}}^{-1}\hat{\underline{\mu}}_{MLE} \right] = \underline{0}$$

$$\underline{\underline{\Sigma}} \left[\underline{\underline{\Sigma}}^{-1} \left(\sum_i \underline{x}_i \right) \right] = \underline{\underline{\Sigma}} \left[n \underline{\underline{\Sigma}}^{-1} \hat{\underline{\mu}}_{MLE} \right]$$

$$\frac{1}{n}\sum_i \underline{x}_i = \hat{\underline{\mu}}_{MLE}$$

MLE in a Bayes (min. error) classifier

Bayes min. error dec. rule:

(4) $$p(\underline{x}|S_k)\,P(S_k) > p(\underline{x}|S_\ell)\,P(S_\ell) \quad \forall\, \ell \neq k \Rightarrow \underline{x} \in S_k$$

(5) Given: $p(\underline{x}|S_k) = f_k(\underline{x}, \underline{\theta}_k)$

($\underline{\theta}_k$: unknown; f_k: known)

Use: $\underline{\theta}_k \approx \hat{\underline{\theta}}^{(k)}_{MLE}$, so that $\hat{p}(\underline{x}|S_k) = f_k(\underline{x}, \hat{\underline{\theta}}^{(k)}_{MLE})$.

Bayes min. error decision rule becomes:

$$\hat{p}(\underline{x}|S_k)\,P(S_k) > \hat{p}(\underline{x}|S_\ell)\,P(S_\ell) \quad \forall\, k \neq \ell \Rightarrow \underline{x} \in S_k$$

(6) with $\hat{p}(\underline{x}|S_j) = f_j(\underline{x}, \hat{\underline{\theta}}^{(j)}_{MLE}) \quad \forall j$.

Can also use $P(S_j) \approx \hat{P}(S_j)$ if $P(S_j)$ are unknown

Comments

1. $\hat{p}(\underline{x} \mid S_k)$ is obtained by plugging in the estimate (MLE) of $\underline{\theta}$ into the assumed or known $p(\underline{x} \mid S_k) = f_k(\underline{x}, \underline{\theta}_k)$

2. Linear Bayes and quadratic Bayes usually mean Bayes (min. error) classifier with assumed normal densities for $p(\underline{x} \mid S_j)$, with $\underline{m}_j$ and $\underline{\underline{\Sigma}}_j$ estimated from the training data using (usually) maximum likelihood estimates.
   It's linear if $\underline{\underline{\Sigma}}_i = \underline{\underline{\Sigma}}_j \; \forall i,j$, and generally quadratic otherwise.

3. Also, linear discriminant analysis (LDA) and quadratic discrimant analysis (QDA) often have similar meaning as linear Bayes and quadratic Bayes
   - except LDA and QDA are often also used as a method of feature-space transformation for dimensionality reduction (similar to FLD and MDA). (The "analysis" term)

View 2: $\underline{\theta}$ is random

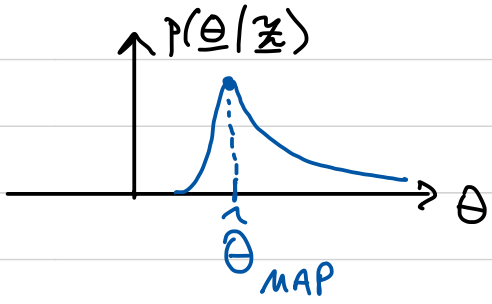
$\underline{\theta}$ has underlying pdf: $p(\underline{\theta})$.
→ Bayesian estimation.

Maximum a-Posteriori (MAP) estimate

(7) $$\hat{\underline{\theta}}_{MAP} \triangleq \underset{\underline{\theta}}{\text{argmax}} \left\{ p(\underline{\theta} | \underline{z}) \right\} = \underset{\underline{\theta}}{\text{argmax}} \left\{ \ln p(\underline{\theta} | \underline{z}) \right\}$$

$p(\underline{\theta} | \underline{z})$ = posterior density of $\underline{\theta}$.

$\hat{\underline{\theta}}_{MAP}$ is the $\underline{\theta}$ that is most probable, given the data $\underline{z}$ (mode of $p(\underline{\theta} | \underline{z})$).



Write $p(\underline{\theta} | \underline{z})$ in terms of $p(\underline{z} | \underline{\theta})$:

(8) $$\begin{cases} p(\underline{\theta} | \underline{z}) = \dfrac{p(\underline{z} | \underline{\theta})\, p(\underline{\theta})}{p(\underline{z})} \\ \ln p(\underline{\theta} | \underline{z}) = \ln p(\underline{z} | \underline{\theta}) + \ln p(\underline{\theta}) - \underbrace{\ln p(\underline{z})}_{\text{const. of } \underline{\theta}} \end{cases}$$

(9) $$\hat{\underline{\theta}}_{MAP} = \underset{\underline{\theta}}{\text{argmax}} \left\{ \underbrace{\ln p(\underline{z} | \underline{\theta})}_{\substack{\text{known} \\ \text{or assumed} \\ (MLE)}} + \underbrace{\ln p(\underline{\theta})}_{\substack{\text{prior} \\ \text{on } \underline{\theta}}} \right\}$$

Prior on $\underline{\theta}$

→ Allows us to insert our prior knowledge of the problem.

Ex: Classify tumors: S_1 = malignant; S_2 = benign.
Assume you normalize or standardize the data so that

$$0 \le x_j \le 1 \ \forall j \quad \text{or} \quad \sigma_j = 1 \ \forall j.$$

We have D features. Let $\underline{\theta} = \underline{w}$.
We will use a linear model so $g(\underline{x}) = \underline{w}^T \underline{x}$ with $g(\underline{x}) \gtrless 0$ (Γ_1 / Γ_2).
Obtain priors from domain experts: a panel of radiologists.

Q: On a scale of 0 to 10, how significant is tumor size in deciding malignancy? (0 = insignifant (not relevant), 10 = very important in deciding malignancy)
A: 8

Q: Does a larger tumor size usually mean a higher or lower chance it is malignant?
A: higher

$\Rightarrow w_1 = +8$

Q: Same questions for tumor symmetry?
A: 5, and more symmetry means less likely to be malignant.

$\Rightarrow w_2 = -5$

Q: Overall, how certain are you of your numeric answers? Can you give a numeric range – for example a significance of 5 could be a range of 4-6, or 3-7, or?
A: 3-7

$\Rightarrow$ Prior $p(\underline{w}) = N(\underline{w} \mid \underline{m}_w, \underline{\underline{\Sigma}}_w)$, $\underline{\underline{\Sigma}}_w = \sigma_w^2 \underline{\underline{I}}$ (assumptions)
$(\underline{m}_w)_1 = 8$, $(\underline{m}_w)_2 = -5$, $\underline{\underline{\Sigma}}_w = 2^2 \underline{\underline{I}}$

Bayesian estimation in a Bayes (min. error) classifier

Bayes min. error dec. rule:

(4) $p(\underline{x}|S_k)P(S_k) > p(\underline{x}|S_\ell)P(S_\ell) \quad \forall \ell \neq k \Rightarrow \underline{x} \in S_k$

(1) Given: $p(\underline{x}|S_k) = f_k(\underline{x}, \underline{\theta}_k) = p(\underline{x}|\underline{\theta}_k, S_k)$

(f_k: known; $\underline{\theta}_k$: unknown)

Two approaches

1. MAP as a point estimate for $\underline{\theta}_k$

$$\underline{\theta}_k \approx \hat{\underline{\theta}}_{MAP}^{(k)}$$

(4) becomes:

$$\hat{p}(\underline{x}|S_k)P(S_k) > \hat{p}(\underline{x}|S_\ell)P(S_\ell) \quad \forall \ell \neq k \Rightarrow \underline{x} \in S_k$$

(10) with $\hat{p}(\underline{x}|S_j) = f_j(\underline{x}, \hat{\underline{\theta}}_{MAP}^{(j)}) \quad \forall j.$

Can also use $P(S_j) \approx \hat{P}(S_j)$ if $P(S_j)$ are unknown

2. Integrate over parameter posterior

Use posterior density of $\underline{\theta}$, $p(\underline{\theta} \mid \underline{\mathcal{Z}})$

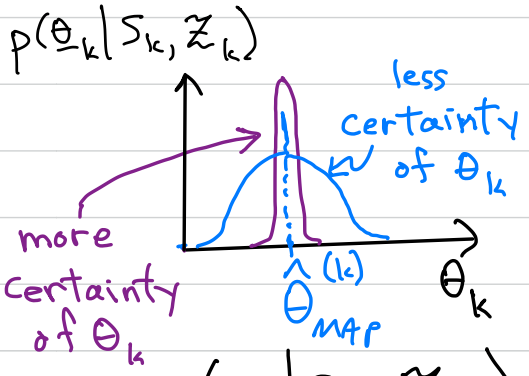
Integrate over $\underline{\theta}$, instead of setting $\underline{\theta} \approx \hat{\underline{\theta}}_{MAP}$

→ Full Bayesian estimate for $p(\underline{x} \mid S_k)$

$$(11) \quad \boxed{p(\underline{x} \mid S_k, \underline{\mathcal{Z}}_k) = \int p(\underline{x} \mid \underline{\theta}_k, S_k, \cancel{\underline{\mathcal{Z}}_k}) p(\underline{\theta}_k \mid S_k, \underline{\mathcal{Z}}_k) \, d\underline{\theta}_k}$$

(given $\underline{\theta}_k$, don't need $\underline{\mathcal{Z}}_k$)

posterior density of $\underline{\theta}_k$ for class S_k



our given or assumed $p(\underline{x} \mid \underline{\theta}_k, S_k)$

from (8) for each class S_k

$p(\underline{x} \mid S_k, \underline{\mathcal{Z}}_k)$ = average likelihood of $S_k, \underline{\mathcal{Z}}_k$

(4) becomes:

$$\hat{p}(\underline{x} \mid S_k) P(S_k) > \hat{p}(\underline{x} \mid S_l) P(S_l) \quad \forall l \neq k \Rightarrow \underline{x} \in S_k$$

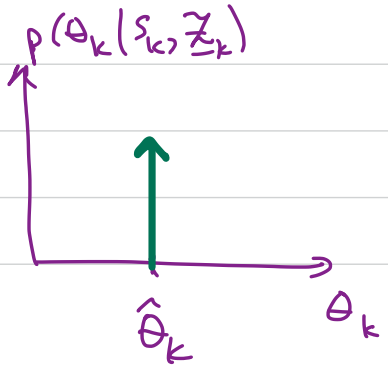
$$(12) \quad \text{with } \hat{p}(\underline{x} \mid S_j) = p(\underline{x} \mid S_j, \underline{\mathcal{Z}}_j) \quad \forall j.$$

Can also use $P(S_j) \approx \hat{P}(S_j)$ if $P(S_j)$ are unknown

Note: if $p(\theta_k \mid S_k, \underline{\mathcal{Z}}_k) = \delta(\theta_k - \hat{\theta}_k)$

$\Rightarrow p(\underline{x} \mid S_k, \underline{\mathcal{Z}}_k) = p(\underline{x} \mid S_k, \hat{\underline{\theta}}_k)$.

(same as approach 1).



Degrees of freedom and constraints in parameter estimation

Ex1: estimate $\underline{m}_k$ only. $(k=1,\cdots,C)$. $\Rightarrow$ D d.o.f. for each class $(p(\underline{x}|S_k), k=1,\cdots,C)$.

constraints = N_k for class S_k $\Rightarrow$ CD total.

N total

Ex2: estimate $\underline{\underline{\Sigma}}_k$ using sample covariance matrix.

$$\hat{\underline{\underline{\Sigma}}}_k = \frac{1}{N_k} \sum_{i=1}^{N_k} \underbrace{(\underline{x}_i - \hat{\underline{m}}_k)(\underline{x}_i - \hat{\underline{m}}_k)^T}_{\underline{\underline{A}}_i}$$

Each matrix $\underline{\underline{A}}_i$ is rank 1.

$\sum_i$ will have rank $\leq D$

what if $N < D$?

the max. rank of $\hat{\underline{\underline{\Sigma}}}$ is $N_k - 1$.

$\Rightarrow$ rank of $\hat{\underline{\underline{\Sigma}}}_k \leq \min\{N_k - 1, D\}$.

$$(\#\text{d.o.f.})_k \le D\,(\text{rank of } \hat{\Sigma}_k) \le D \min\{N_k - 1, D\}.$$

$$\text{Total: } D \sum_{k=1}^{C} \min\{N_k - 1, D\}. \le CD^2.$$

$$\#\text{constraints} = N. \longleftarrow$$

Est. m_k and $\hat{\Sigma}_k$, $k = 1, \cdots, C \Rightarrow$ d.o.f. $= CD^2 + CD = \underline{CD(D+1)}$

(dens.-est., histogr. method: M^D).
→ C classes $\Rightarrow$ $C \cdot M^D$