MA8701 Advanced methods in statistical inference and learning

L2: Classification and statistical decision theory, model selection and assessment

Mette Langaas

~~1/12/23~~

13.01.2023

notes in class

with some additions after class

Plan

Continue with the decision theoretic framework from L1, but now for classification. Bias-variance trade-off.

- Classification - should not be new (ESL Ch 4.1-4.5, except 4.4.4)
- Statistical decision theoretic framework for classification (ESL 2.4)
- and the bias-variance trade-off (ESL 2.9 and 7.2-7.3)

Decision theoretic framework — now classification

$\underline{X} \in \mathbb{R}^p$;

$G \in \mathcal{G} = \{1, 2, \ldots, K\}$ or $K = 2$ give $\mathcal{G} = \{0, 1\}$

$\hat{G}(\underline{X})$: our function of interest – to predict G

What is the optimal choice?

$L(G, \hat{G}(\underline{X}))$: loss function – assign numerical value

$\mathbb{L}$ $K \times K$

with "general" values of diagonal

$$
\begin{array}{c|ccccc}
 & \multicolumn{5}{c}{\hat{G}(\underline{X})} \\
G & 1 & 2 & 3 & \cdots & K \\
\hline
1 & 0 & 1 & 1 & \cdots & 1 \\
2 & 1 & 0 & & & \\
\vdots & \vdots & & \ddots & & \\
K & 1 & & & & 0
\end{array}
\qquad \mathbb{1}\mathbb{1}^T - I
$$

if all loss = 1 for misclassification

Find $\hat{G}(X)$ to minimize EPE

prediction

expected

error

$p(x,G) = p(G|x) \cdot p(x)$

$$\underset{G,X}{E}\left[L(G,\hat{G}(X)) \right] = \underset{X}{E}\left[\underset{G|X}{E} \left(L(G,\hat{G}(X)) \right) \right]$$

$$= E_x \left\{ \sum_{k=1}^{K} L(g_k, \hat{G}(X)) \, P(G=g_k | X=x) \right\}$$

one element of the $\mathbb{L}$ matrix

For each $X=x$ we have

$$\hat{G}(x) = \underset{g \in \mathcal{G}}{\operatorname{argmin}} \sum_{k=1}^{K} L(g_k, \hat{G}(x)) \, P(G=g_k | X=x)$$

the loss of saying $\hat{G}(X)$ when g_k is true and if 0-1 loss is used, and let g be the true class

$$\sum_{k=1}^{K} L(g_k, \hat{G}(x)) \cdot P(G=g_k \mid X=x)$$

$$= 0 \cdot P(G=g \mid X=x) + \underbrace{\sum_{g_k \neq g} 1 \cdot P(G=g_k \mid X=x)}_{1-P(G=g \mid X=x)}$$

$$\hat{G}(X) = \underset{g \in \mathcal{G}}{\arg\min}\,(1-P(G=g \mid X=x)) = \underset{g \in \mathcal{G}}{\arg\max}\, P(G=g \mid X=x)$$

classify to the most probable class

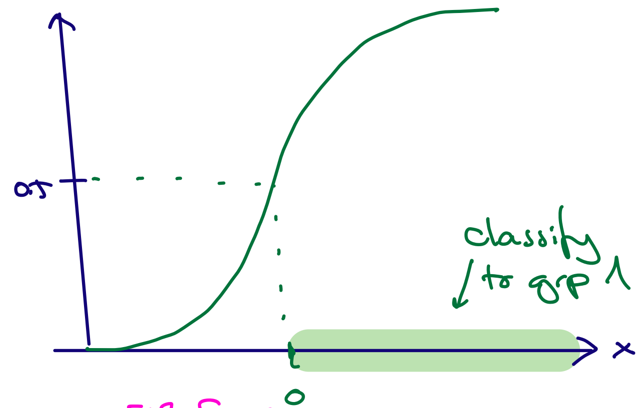
$\Rightarrow$ our optimal classifier is called the Bayes classifier

Can we calculate such a classifier?

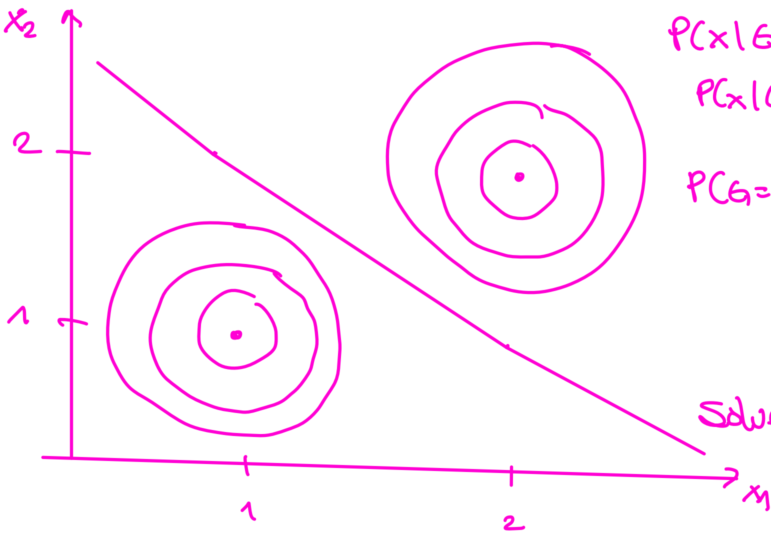
Yes! If we know $P(G=g \mid X=x)$ for all g and x

Ex 1: $p=1,\ K=2$

$$P(G=1 \mid X=x) = \frac{\exp(0+0.8x)}{1+\exp(0+0.8x)}$$



Ex 2: $P(G=0) = P(G=1) = \frac{1}{2}$



$$P(x \mid G=0) = N\left(\begin{bmatrix}1\\1\end{bmatrix}, \begin{bmatrix}1&0\\0&1\end{bmatrix}\right)$$

$$P(x \mid G=1) = N\left(\begin{bmatrix}2\\2\end{bmatrix}, \begin{bmatrix}1&0\\0&1\end{bmatrix}\right)$$

$$P(G=1 \mid X=x) = \frac{P(x \mid G=1)\cdot P(G=1)}{P(x \mid G=1)\cdot P(G=1) + P(x \mid G=0)\cdot P(G=0)}$$

$$P(G=0 \mid X=x) = 1 - P(G=1 \mid X=x)$$

Solve $P(G=0 \mid X=x) = P(G=1 \mid X=x)$ to find boundary $\rightarrow$ See Exerc. for mixture!

Group discussion

1) What do we know about classification? (TMA4268 and TMA4315 mainly, or ESL ch 4.1-4.5, except 4.4.4)
   - What is the difference between discrimination and classification?
   - What are the sampling vs diagnostic paradigm? Give an example of one method of each type.
   - Give an example of one parametric and one non-parametric classification method.
2) Logistic regression is by many seen as the "most important method in machine learning". What do we remember about logistic regression? (Will be a very important method in Part 2.)
3) What "changes" need to be done to 2) when we have $K > 2$ classes?

1) a) discrimination : focus on understanding how groups can be separated.

classification : predict class memberships

b) sampling : model $p(g)$ and $p(x|g)$: LDA, QDA

diagnostic : model $p(g|x)$: kNN, SVM, logistic regression)

c) parametric : logistic regression, LDA

non-parametric : k-NN

trees - random forests

can be written with formulas

but, will soon be large

expressions

2) LOGISTIC REGRESSION ← GLM, binary response and logit link

a) 1) $Y_i \sim bin(1, \pi_i)$ (also possible to use (n_i, π_i) if we consider "covariate patterns" - and we need that for e.g. deviance statistics)

2) $\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi}$

3) $\underset{\substack{\mu_i \\ \pi_i}}{E(Y_i)} = \dfrac{e^{\eta_i}}{1+e^{\eta_i}}$ response function

$\underset{\substack{\uparrow \\ g(\mu_i)\text{ link function}}}{\eta_i} = \text{logit}(\pi_i) = \log\left(\dfrac{\pi_i}{1-\pi_i}\right)$

b) $f(y_i; \pi_i) = \underbrace{\binom{1}{y_i}}_{1} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$

In class: $y_i = 0 \Rightarrow \binom{1}{0} = 1$
$y_i = 1 \Rightarrow \binom{1}{1} = 1$

$$L(y_1, \dots, y_N) = \prod_{i=1}^{N} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

$$l(\beta) = \ln L(y_1, \dots, y_N) = \sum_{i=1}^{N}\left(y_i \ln \pi_i + (1-y_i)\ln(1-\pi_i)\right)$$

$$= \sum_{i=1}^{N}\left(y_i \ln \pi_i + \ln(1-\pi_i) - y_i \ln(1-\pi_i)\right) = \sum_{i=1}^{N}\left(y_i \cdot \ln\left(\frac{\pi_i}{1-\pi_i}\right) + \ln(1-\pi_i)\right)$$

REMARK: as a function of β — add connection $\pi_i \leftrightarrow \mu_i \leftrightarrow \beta$

$$l(\beta) = \sum_{i=1}^{N} \left\{ y_i x_i^T \beta - \ln\left(1 + \exp\left(x_i^T \beta\right)\right) \right\}$$

$$U(\beta) = \frac{\partial l}{\partial \beta}$$

$(p+1)$ vector

REMARK: $U(\beta) = \sum_{i=1}^{N} U_i(\beta)$

$$U_i(\beta) = \frac{\partial l_i(\beta)}{\partial \beta} = \dots = x_i\left(y_i - \frac{\exp(x_i^T\beta)}{1+\exp(x_i^T\beta)}\right)$$

For "ALL" GLM's $U(\beta) = X^T D \Sigma^{-1} (y - \mu)$

X $n \times p$

$D = \operatorname{diag}(h'(\eta_i))$

$\Sigma = \operatorname{diag}(\operatorname{Var}(Y_i))$

y $n \times 1$

$\mu = E(Y) = h(\eta)$ $n \times 1$

observed Fisher info
(minus Hessian)

$$H(\beta) = -\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T} = \dots = \sum_{i=1}^{n} x_i x_i^T \pi_i (1 - \pi_i) \qquad \pi_i = \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)}$$

$(p+1) \times (p+1)$

$\underbrace{E(H(\beta))}_{I(\beta)} = H(\beta)$ in this case: general property for GLM with canonical link

$I(\beta) \leftarrow$ Expected Fisher information matrix

c) Finding $\hat{\beta}$ using Newton-Raphson: $U(\beta) \approx U(\beta^{(1)}) + H(\beta^{(1)})(\beta - \beta^{(1)}) = 0$

$\uparrow$ solve $U(\beta) = 0$

$(p+1) \times 1$ $\qquad$ $(p+1) \cdot (p+1)$

$$\beta^{(r+1)} = \beta^{(r)} + H(\beta^{(r)})^{-1} U(\beta^{(r)})$$

use $I(\beta)$ = Fisher scoring
and here $H(\beta) = I(\beta)$ so the same

d) Asymptotic properties

$$\hat{\beta} \approx N_{p+1}(\beta, I^{-1}(\beta))$$

Slutsky: replace with $\hat{\beta}$ ok
(TMA4295)

NB: For part 2 on GLM + lasso/ridge we need a bit more on the estimation process!

We have lectures Monday 10-12 and Friday 10-12.

Is that enough? What if you have questions? Google and fellow students to answer? Or would you like to have

- digital mattelab - to ask questions and everyone helps to answer?
- office hours to ask about theory/exercises/etc or
- booked S21 or 656 to work together and ask questions?

⇒ METTE will have office hours in 1236 before class

Mondays 9–10

Fridays 9–10

and can also be contacted by email to book session!

(and this may change during the semester if needed)

MODEL ASSESSMENT AND SELECTION

ESL 7.1-7.6, 7.10-7.12
(L2-3-4)

$\hat{f}(X)$: estimated from training data

ASSESSMENT

evaluate the performance of a selected model

↓

use test data and loss function

SELECTION

select the best model

↓

use validation data and loss function

DATA RICH SITUATION

⋮

What if we only have training data?

Can we perform model selection and assessment?

Plan

L2
1) Look at EPE(x_0) (now called Err(x_0) after we have estimated f) and how model complexity can be broken down into irreducible error, squared bias and variance (should be known from before)
2) Study EPE (Err) unconditional and conditional on the training set

L3
3) Study optimism of the training error rate, and how in-sample error may shed light on methods for model selection (like AIC, Mallows Cp)

L4
4) Cross-validation and .632 bootstrap estimates of EPE
5) How will we build on this in Parts 2-4?

BIAS VARIANCE TRADE-OFF

Additive regression model $\quad Y = f(X) + \varepsilon \quad$ $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma_\varepsilon^2$ $\quad$ AND use quadratic loss

Have estimated $\hat{f}(X)$ from training data $\mathcal{T}$

EPE is from now on called Err and

$$Err(\hat{f}) = \mathop{E}_{XY}\left[\left(Y - \hat{f}(X) \right)^2 \right] = \mathop{E}_{x_0}\Big[\underbrace{Err(x_0)}_{E\left((Y - \hat{f}(X))^2 \mid X = x_0 \right)} \Big]$$

irreducible error

varience at prediction

bias at prediction

truth $f(x_0)$

$$Err(x_0) = \sigma_\varepsilon^2 + Var(\hat{f}(x_0)) + Bias(\hat{f}(x_0))^2$$

$$Bias(\hat{f}(x_0))^2 = (E(\hat{f}(x_0)) - f(x_0))^2$$

Group activity

- Remind yourself on how this derivation was done and the meaning of each term.
- What is the role of x_0 here? How can you get rid of x_0?

Err – but now conditional on training set

(now: general L)

τ: training set

$$\mathrm{Err}_\tau = E\left(L(Y, \hat{f}(X)) \mid \tau \right)$$

$$\mathrm{Err} = E_\tau\left(\mathrm{Err}_\tau \right)$$

$$\overline{\mathrm{err}} = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{f}(x_i))$$

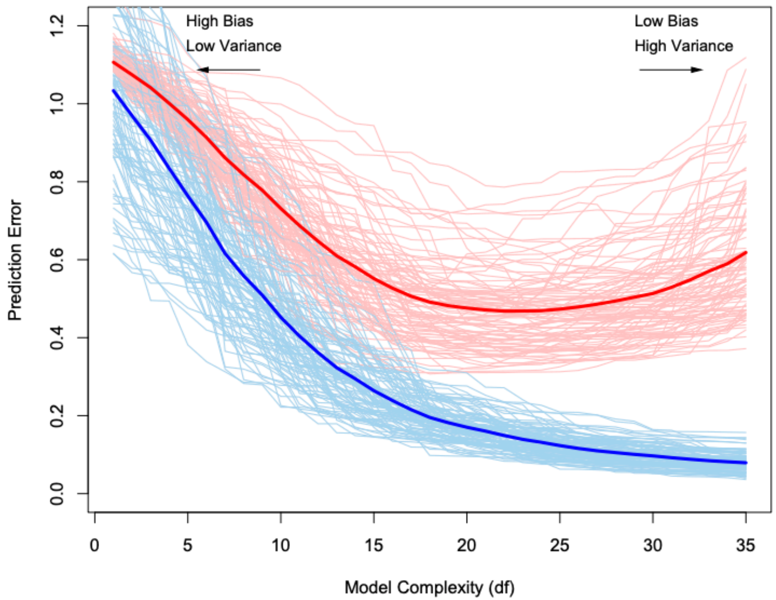
$(X, Y) \rightarrow x^1, y^1$ in τ – to estimate $\hat{f}$

$(X, Y) \rightarrow x^0, y^0$ new test data to evaluate $\hat{f}$

Group discussion

Look at Figure 7.1 (with figure caption) on page 220 in the ESL book. The text reads that "100 simulated training sets of size 50" and that "lasso produced sequence of fits" (this means that we have different model complexities on the x-axis).
Explain what you see - in particular what are the red and blue lines and the bold lines. What can you conclude from the figure?

- Red lines= $\mathrm{Err}_{\mathcal{T}}$ estimate (maybe just one test set?)
- Bold red line= $\mathrm{Err} = E_{\mathcal{T}}(\mathrm{Err}_{\mathcal{T}})$ estimate
- Blue lines= $\frac{1}{100}\sum_{j=1}^{100} \overline{\mathrm{err}}_j$ mean of training error over the 100 training sets
- Bold blue line= $\overline{\mathrm{err}}_1, \ldots, \overline{\mathrm{err}}_{100}$



100 simulated training sets of size 50, lasso produced sequence of fits. What do you see? What are the red and blue lines? Conclusions?

Read for yourself!

Loss function and training error for classification

- $X \in \Re^p$
- $G \in G = \{1, \ldots, K\}$
- $\hat{G}(X) \in G = \{1, \ldots, K\}$

0-1 loss with $\hat{G}(X) = \text{argmax}_k \hat{p}_k(X)$

$$L(G, \hat{G}(X)) = I(G \neq \hat{G}(X))$$

-2-loglikelihood loss (why -2?):

$$L(G, \hat{p}(X)) = -2\text{log}\hat{p}_G(X)$$

Test error (only replace $\hat{f}$ with $\hat{G}$):

$$\mathsf{Err}_T = \mathsf{E}[L(Y, \hat{G}(X)) \mid T]$$

$$\mathsf{Err} = \mathsf{E}[\mathsf{E}[L(Y, \hat{G}(X)) \mid T]] = \mathsf{E}[\mathsf{Err}_T]$$

Training error (for 0-1 loss)

$$\overline{\mathsf{err}} = \frac{1}{N} \sum_{i=1}^{N} I(g_i \neq \hat{g}(x_i))$$

Training error (for -2loglikelihood loss)

$$\overline{\mathsf{err}} = -\frac{2}{N} \sum_{i=1}^{N} \mathsf{log} \hat{p}_{g_i}(x_i)$$

Exercises

REMEMBER TO DO THIS!

What are the most important results from the "Statistical decision theoretic framework"?

- What are results to remember for regression and for classification?
- How would you use these results?

Look into the derivation for the bias and variance decomposition

- for kNN in Equation 7.10 and
- OLS in Equation 7.11 on pages 222-223 of ESL.

Bayes classier, Bayes decision boundary and Bayes error rate

Solve TMA4268 exam problem 9 in 2019 at https://www.math.ntnu.no/emner/TMA4268/Exam/V2019e.pdf

Key results from logistic regression

a) What are the three components of a generalized linear model?

b) What are these three for a logistic regression?

c) Parameter estimation

How are regression parameters estimated for the GLM, and for logistic regression in particular?
Does it matter if you use the observed or expected information matrix for logistic regression?

d) Asymptotic distribution

What is the asymptotic distribution of the estimator for the regression parameter $\widehat{\beta}$? How can that be used to construct confidence intervals or perform hypothesis tests?

e) Deviance

How is the deviance defined in general, and how is this done for

Next week (week 2) : we finish ESL chapter 7, and if time continue to missing data (or that is week 3)