Statistical Learning IV (19-11-2015)

11.15 - 13.00, 13.45 - 15.30

1. Probability Theory
2. Bayesian Statistics
   a] Intro
   b] Example: Laplace rule of succession
   c] MAP interpretation of Ridge and Lasso

Classification:

3. Problems with Least Squares for Classification ← skipped
5. Linear Discriminant Analysis
6. Naive Bayes
4. Plug-in estimators

1. Probability Theory

Bayes' Rule:

$$Pr(A|B) = \frac{Pr(B|A) \cdot Pr(A)}{Pr(B)}$$

often used in Bayesian <u>and</u> standard frequentist statistics

~~[illegible]~~ <u>Densities are slippery:</u>

* Location of maximum depends on choice of parametrisation
* Uniform density in one parametrisation is not uniform in another parametrisation

(see slides)

2. Bayesian Statistics

a] Intro

Probability model: $\mathcal{P} = \{ P_\theta(X, Y) \mid \theta \in \Theta \}$ or $\mathcal{P} = \{ P_\theta(Y) \mid \theta \in \Theta \}$

Frequentist statistics (standard):

- optimal parameter θ^* is fixed, but unknown
- diversity of methods
- need proofs/experiments to justify methods

Bayesian statistics:

- pretend that true parameter θ^* is a random variable distributed according to prior distribution $\pi(\theta)$ that we know.

$T = Y_1, \ldots, Y_N$ (no features for simplicity)

$$\Pr(T, \theta) = P_\theta(T) \cdot \pi(\theta)$$ is joint distribution of data and parameters

Since we know the full joint distribution, can simply use probability theory to compute any probability we are interested in. ← single method

↳ if we estimate different probabilities this way, they are beautifully consistent.

~~But~~ Bayesian premise is too strong:

- prior π is chosen for computational or information theoretic properties in practice, so cannot ~~blind~~ blindly assume θ^* is random sample from π.
- need proofs/experiments to justify Bayesian methods

Modern frequentist motivation:

- If we choose π right, then often works really well (both in theory and in practice).
- Learn faster if true θ^* has high prior probability, slower if true θ^* has small prior probability, so can use π to express <u>prior knowledge</u> about our data.

Posterior Distribution:

$$\pi(\theta|T) := Pr(\theta|T) = \frac{Pr(\theta, T)}{Pr(T)} = \frac{P_\theta(T) \cdot \pi(\theta)}{Pr(T)}$$

- Often puts its probability mass closer and closer to θ^* as $N \to \infty$. (vdVaart et al.)
- Expresses uncertainty about θ

Predictive Distribution:

$$Pr(Y|T) = \int_\Theta P_\theta(Y) \cdot \pi(\theta|T)\, d\theta = \frac{Pr(Y,T)}{Pr(T)}$$

new sample outside of training set

often better predictions than frequentist $P_{\hat{\theta}}(Y)$ because $\pi(\theta|T)$ keeps track of uncertainty better than fixed single choice $\hat{\theta}$.

b] Example: Laplace Rule of Succession

Bernoulli model: $P_\theta(Y) = \begin{cases} \theta & \text{for } Y=1 \\ 1-\theta & \text{for } Y=0 \end{cases}$ $\quad \theta \in [0,1]$

Maximum likelihood: $\hat{\theta} = \frac{n_1}{N} \rightarrow P_{\hat{\theta}}(Y=1) = \hat{\theta} = \frac{n_1}{N}$ ← dangerous for prediction if $n_1 = 0$

Suppose $\pi(\theta) = 1$ is uniform prior density, ⇐ not the same as "no prior knowledge" because depends on parametrisation

Then $Pr(Y=1|T) = \frac{n_1+1}{N+2}$

(Laplace, 1814)

$Pr(Y=0|T) = \frac{n_0+1}{N+2}$

beta function

Proof: $\frac{Pr(Y,T)}{Pr(T)} = \frac{\int P_\theta(Y,T) \cdot \pi(\theta)\, d\theta}{\int P_\theta(T) \cdot \pi(\theta)\, d\theta} = \frac{\int \theta^{n_1+Y}(1-\theta)^{n_0+1-Y} d\theta}{\int \theta^{n_1}(1-\theta)^{n_0} d\theta} = \frac{n_Y+1}{N+2}$ □

c] MAP interpretation of Ridge and Lasso

Maximum a Posteriori (MAP) parameters:

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\ \pi(\theta|T) = \underset{\theta}{\operatorname{argmax}}\ \frac{P_\theta(T)\cdot\pi(\theta)}{Pr(T)}$$

$$= \underset{\theta}{\operatorname{argmax}}\ P_\theta(T)\cdot\pi(\theta)$$

- maximizes posterior density, so for continuous parameters depends on parametrisation!
- "real" Bayesians prefer prediction with predictive distribution.

Ridge / Lasso:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\ RSS(\beta) + \lambda\, pen(\beta)$$

$$\text{ridge: } pen(\beta) = \sum_{j=2}^{p} \beta_j^2 \qquad \text{lasso: } pen(\beta) = \sum_{j=2}^{p} |\beta_j|$$

Suppose Gaussian noise:

$$Y = X^T\beta + \varepsilon \qquad \varepsilon \sim \mathcal{N}(0,\sigma^2)$$

$$P_\beta(Y_1,\dots,Y_N|X_1,\dots,X_N) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(Y_i - X_i^T\beta)^2}{2\sigma^2}}$$

$$\hat{\beta}_{MAP} = \underset{\beta}{\operatorname{argmax}}\ P_\beta(Y_1,\dots,Y_N|X_1,\dots,X_N)\cdot\pi(\beta)$$

$$= \underset{\beta}{\operatorname{argmin}}\ -\log P_\beta(Y_1,\dots,Y_N|X_1,\dots,X_N) - \log\pi(\beta)$$

$$= \underset{\beta}{\operatorname{argmin}}\ N\cdot\left(-\log\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \frac{1}{2\sigma^2}RSS(\beta) - \log\pi(\beta)$$

$$= \underset{\beta}{\operatorname{argmin}}\ RSS(\beta) - 2\sigma^2\log\pi(\beta)$$

Suppose data have been pre-processed such that we can assume that the intercept $\hat{\beta}_1 = 0$.

Ridge: Choose π s.t.

$$\beta_1 = 0 \quad \text{with prob. } 1$$
$$(\beta_2, \ldots, \beta_p) \sim \mathcal{N}(0, \tau^2 I)$$

$$-\log \pi(\beta) = \sum_{j=2}^{p} -\log\left(\frac{1}{\sqrt{2\pi\tau^2}} \cdot e^{-\frac{(\beta_j - 0)^2}{2\tau^2}}\right)$$

$$= (p-1)\left(-\log \frac{1}{\sqrt{2\pi\tau^2}}\right) + \sum_{j=2}^{p} \frac{\beta_j^2}{2\tau^2}$$

$$\hat{\beta}_{MAP} = \underset{\beta}{\operatorname{argmin}}\ RSS(\beta) + \frac{\sigma^2}{\tau^2} \sum_{j=2}^{p} \beta_j^2$$

is ridge with $\lambda = \frac{\sigma^2}{\tau^2}$

Is also posterior mean $\mathbb{E}_{\pi(\beta \mid T)}[\beta]$

Lasso: Choose π s.t.

$$\beta_1 = 0 \quad \text{with prob. } 1.$$
$$(\beta_2, \ldots, \beta_p) \sim \prod_{j=2}^{p} \frac{1}{2\tau} e^{-\frac{|\beta_j|}{\tau}}$$

$$-\log \pi(\beta) = \sum_{j=2}^{p} -\log\left(\frac{1}{2\tau} \cdot e^{-\frac{|\beta_j|}{\tau}}\right)$$

$$= (p-1)\left(-\log \frac{1}{2\tau}\right) + \sum_{j=2}^{p} \frac{|\beta_j|}{\tau}$$

$$\hat{\beta}_{MAP} = \underset{\beta}{\operatorname{argmin}}\ RSS(\beta) + \frac{2\sigma^2}{\tau} \sum_{j=2}^{p} |\beta_j|$$

is Lasso with $\lambda = \frac{2\sigma^2}{\tau}$.

Remarks:

- "real" Bayesian prefers predicting with predictive distribution
- MAP + CV to determine λ is very unBayesian.

(see "Classification" at bottom first)

3. Problems with Least Squares for Classification

For K classes:

$$Y_i^k = \begin{cases} 1 & \text{if } Y_i = k \\ 0 & \text{if } Y_i \neq k \end{cases} \qquad \text{for } k = 1, \ldots, K$$

X_i : features

Bayes optimal: $\arg\max_k \Pr(Y=k \mid X)$

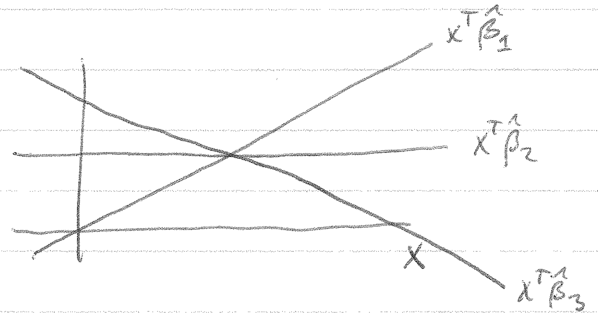
Idea: - estimate $\Pr(Y=k \mid X)$ by $x^T\hat{\beta}_k$
where $\hat{\beta}_k$ is least squares estimate for responses $Y_1^k, \ldots, Y_N^k$ and features $X_1, \ldots, X_N$.

- classify by $\arg\max_k x^T\hat{\beta}_k$

Problems:

* probabilities ~~often~~ usually do not behave linearly.
  E.g. $x^T\hat{\beta}_k$ can be negative or larger than 1.

* masking: when $K \geq 3$ and p small:



← class 2 never selected.

Figures 4.2, 4.3 show how this can happen.

Classification: e.g. spam filtering, digit recognition

$$T = \begin{pmatrix} G_1 \\ x_1 \end{pmatrix}, \ldots, \begin{pmatrix} G_N \\ x_N \end{pmatrix} \overset{\text{i.i.d.}}{\sim} p^*$$

e.g. $G \in \{\text{red}, \text{green}, \text{blue}\} \to Y \in \{1, 2, 3\}$

4. Plug-in Estimators

$$p^*(y=k\mid x) = \frac{p^*(x\mid y=k)\cdot p^*(y=k)}{p^*(x)} = \frac{p^*(x\mid y=k)\cdot p^*(y=k)}{\sum_{k'} p^*(x, y=k')}$$

$$= \frac{p^*(x\mid y=k)\cdot p^*(y=k)}{\sum_{k'} p^*(x\mid y=k')\cdot p^*(y=k')} = \frac{f_k(x)\cdot \pi_k}{\sum_{k'} f_{k'}(x)\,\pi_{k'}}$$

Notation: $f_k(x) = p^*(x\mid y=k)$
$\pi_k = p^*(y=k)$

Plug-in estimators:

- estimate $p^*(y=k\mid x)$ by $\hat{p}(y=k\mid x)$ and predict with $\underset{k}{\operatorname{argmax}}\ \hat{p}(y=k\mid x)$
- sufficient to estimate f_k and π_k for all k:

$$\hat{f}(x) = \underset{k}{\operatorname{argmax}}\ \hat{p}(y=k\mid x) = \underset{k}{\operatorname{argmax}}\ \frac{\hat{f}_k(x)\,\hat{\pi}_k}{\sum_{k'} \hat{f}_{k'}(x)\,\hat{\pi}_{k'}}$$

$$= \underset{k}{\operatorname{argmax}}\ \hat{f}_k(x)\,\hat{\pi}_k$$

N.B. Although we have used Bayes' rule, there is nothing Bayesian about these methods!

5. Linear Discriminant Analysis (LDA)

Model: $f_k(x)$ is $\mathcal{N}(\mu_k, \Sigma_k)$ is multivariate Gaussian (see fig. 4.5)

Take $\Sigma_k = \Sigma$ the same for all k

Parameter estimates from T:

$\hat{\pi}_k = \frac{n_k}{N}$ where n_k is number of observations in class k.

$\hat{\mu}_k = \frac{1}{n_k} \sum_{i: y_i = k} x_i$ is mean of x_i in class k

$$\hat{\Sigma} = \frac{1}{N-K} \sum_{k=1}^{K} \sum_{i: y_i = k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$$

$$\begin{aligned}
\arg\max_k \hat{f}_k(x) \cdot \hat{\pi}_k &= \arg\max_k \log \hat{f}_k(x) + \log \hat{\pi}_k \\
&= \arg\max_k \log\left(\frac{1}{(2\pi)^{p/2} |\hat{\Sigma}|^{1/2}} e^{-\frac{1}{2}(x-\hat{\mu}_k)^T \hat{\Sigma}^{-1} (x - \hat{\mu}_k)} \right) + \log \hat{\pi}_k \\
&= \arg\max_k -\tfrac{1}{2}(x - \hat{\mu}_k)^T \hat{\Sigma}^{-1} (x - \hat{\mu}_k) + \log \hat{\pi}_k \\
&= \arg\max_k \cancel{-\tfrac{1}{2} x^T \hat{\Sigma}^{-1} x} + x^T \hat{\Sigma}^{-1} \hat{\mu}_k - \tfrac{1}{2} \hat{\mu}_k^T \hat{\Sigma}^{-1} \hat{\mu}_k + \log \hat{\pi}_k
\end{aligned}$$

linear in x,
so decision boundary between any two classes also linear in x.
Hence name LDA

6. Naive Bayes ← used in spamfilters

Model: features are <u>independent</u>:

$$f_k(x_i) = \prod_{j=1}^{p} f_{kj}(x_{ij})$$

← useful simplification if p is very large, so ~~good exact~~ precise density estimation is impossible

π_k

Continuous features: $x_{ij} \in \mathbb{R}$

Fit Gaussian for each dimension j separately.
(similar to LDA)

Discrete features:

x_i is e-mail message
x_{ij} is word in j-th position

(1) Forget position j of each word "bag of words"

(2) Use separate multinomial model per class k

I.E. m possible words, $x_{ij} \in \{1, \dots, m\}$ ← represent words by their nr. in list of possible words.

parameters $\theta_1^k, \dots, \theta_m^k$ for each class k

$$f_{kj}(x_{ij}) = \theta_{x_{ij}}^k$$

Too simple/naive?

~~Only need $\hat{P}(Y=k|X) > \hat{P}(Y=\ell|X)$ for right class k and wrong class ℓ~~

~~- Possible even if $\hat{P}$ is poor probability estimate~~

- Only need $\hat{P}(Y=k|X) > \hat{P}(Y=\ell|X)$ whenever $P^*(Y=k|X) > P^*(Y=\ell|X)$.
- Possible even if $\hat{P}$ is poor estimate of P^*!