

# ST451 - Lent term

## Bayesian Machine Learning

Kostas Kalogeropoulos

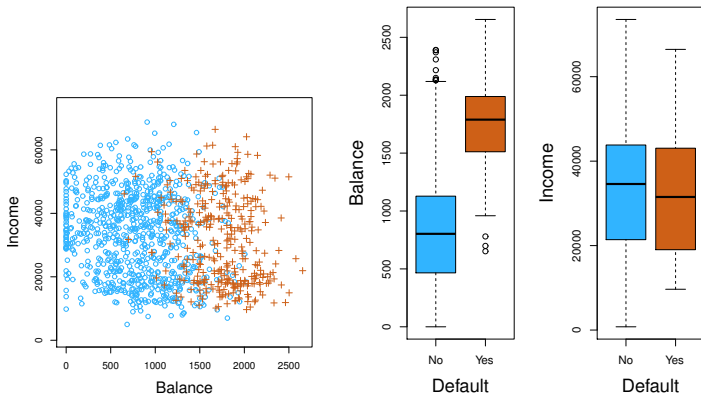
Linear Models for Classification

# Summary of last lecture

- **Bayesian Linear regression:** derivation of posterior when both  $\beta$  and  $\sigma^2$  are unknown.
- Bayesian multi-parameter models, focus on the **marginal** posterior  $\pi(\beta|y, X, \sigma^2)$
- **Bayesian model choice:** Bayes factor, Jeffreys Lindley paradox and unit information priors.
- **Implementation:** Explicit formulae for the posterior mean, Monte Carlo for credible intervals and prediction, 'candidate property' for marginal likelihood calculation.

## Motivating Example

'Default' dataset consist of three variables: annual income, credit card balance and whether or not the person has defaulted in his/her credit card.



The aim is to build a model to predict whether a person will default based on annual income and monthly credit card balance.

# Classification

Generally we will assume that have a number of covariates or features (denoted by  $X$ ) as well as the response  $y$  which is now a **categorical variable** taking values  $c_1, \dots, c_K$ .

Usually we will assume that  $k = 2$  (binary classification) but also consider  $k > 2$  multiple classes

Existing approaches can be split into two categories:

- 1 **Generative models:** specify  $\pi(X|c_k)$ , so that we can *generate*  $X$ , assign prior probabilities on each  $c_k$  and use Bayes theorem to obtain  $\pi(c_k|X)$ . e.g. linear and quadratic discriminant analysis.
- 2 **Discriminative models:** specify the model (likelihood)  $\pi(c_k|X)$  and perform statistical inference and prediction as in linear regression. e.g. logistic and probit regression

# Outline

- 1 Discriminative Models / Logistic Regression
- 2 Bayesian Logistic Regression
- 3 Generative Models
- 4 Assessing prediction in classification

# Outline

- 1 Discriminative Models / Logistic Regression
- 2 Bayesian Logistic Regression
- 3 Generative Models
- 4 Assessing prediction in classification

# Logistic regression

Model for  $(y_i, X_i)$ :

$$y_i = \text{Bernoulli}(\pi(c_k|X_i))$$
$$\pi(c_k|X_i) = \sigma(X_i\beta) \text{ or else } \log\left(\frac{\pi(c_k|X_i)}{1 - \pi(c_k|X_i)}\right) = X_i\beta$$

## Interpretation of coefficients:

- $X$  consists of either dummy or continuous variables.
- A dummy variable  $Z$  is an indicator of a category - say  $A$ . Its  $\beta$  coefficient reflects the **log-odds ratio** between  $A$  and  $A^c$ .

$$\log\left(\frac{\frac{p(y_i=1|X=1)}{1-p(y_i=1|X=1)}}{\frac{p(y_i=1|X=0)}{1-p(y_i=1|X=0)}}\right)$$

- The coefficient of a continuous variable  $X_c$  reflects the **log odds ratio** for a unit change in  $X_c$ .

# Logistic regression

Check your understanding on the following output.

	Coefficient	Std. error	Z-statistic	P-value
<b>Intercept</b>	-10.6513	0.3612	-29.5	<0.0001
<b>balance</b>	0.0055	0.0002	24.9	<0.0001

	Coefficient	Std. error	Z-statistic	P-value
<b>Intercept</b>	-3.5041	0.0707	-49.55	<0.0001
<b>student[Yes]</b>	0.4049	0.1150	3.52	0.0004

	Coefficient	Std. error	Z-statistic	P-value
<b>Intercept</b>	-10.8690	0.4923	-22.08	<0.0001
<b>balance</b>	0.0057	0.0002	24.74	<0.0001
<b>income</b>	0.0030	0.0082	0.37	0.7115
<b>student[Yes]</b>	-0.6468	0.2362	-2.74	0.0062

Note that the coefficient of 'student' is **positive** in table 2 and **negative** in table 3. How can we interpret this?



# Logistic regression - maximum likelihood

The likelihood, log-likelihood, gradient and Hessian can be written as

$$\begin{aligned}f(y|X, \beta) &= \prod_i \left\{ \sigma(X_i\beta)^{y_i} [1 - \sigma(X_i\beta)]^{1-y_i} \right\}. \\ -\ell(\beta) &= -\sum_i \{ y_i \log(\sigma(X_i\beta)) + (1 - y_i) \log(1 - \sigma(X_i\beta)) \}, \\ -\nabla_\beta \ell(\beta) &= -\sum_i \left\{ \frac{y_i \nabla_\beta \sigma(X_i\beta) (1 - \sigma(X_i\beta)) - (1 - y_i) \nabla_\beta \sigma(X_i\beta) \sigma(X_i\beta)}{\sigma(X_i\beta) (1 - \sigma(X_i\beta))} \right\} \\ &\quad \text{using } \nabla_x \sigma(x) = \sigma(x) (1 - \sigma(x)) \text{ gives} \\ &= \sum_i y_i (1 - \sigma(X_i\beta)) X_i^T - (1 - y_i) \sigma(X_i\beta) X_i^T \\ &= \sum_i (\sigma(X_i\beta) - y_i) X_i^T = X^T (\sigma(X\beta) - y) \\ H(\beta) &= \sum_i \sigma(X_i\beta) (1 - \sigma(X_i\beta)) X_i^T X_i = X^T S X,\end{aligned}$$

where  $S$  is a diagonal matrix with entries  $\sigma(X\beta) (1 - \sigma(X\beta))$

# Logistic regression - maximum likelihood

- There is **no closed form** solution but the Newton-Raphson maximisation algorithm can be used given  $\nabla_{\beta}\ell(\beta)$  and  $H_{\beta}$ .

$$\beta_{\text{new}} = \beta_{\text{old}} - H(\beta_{\text{old}})^{-1} \nabla_{\beta}\ell(\beta)|_{\beta=\beta_{\text{old}}}.$$

- Use of normal CDF as a function instead of the sigmoid provides the **probit** regression.
- There is no conjugate prior for  $\beta$  so the posterior is **not available** in closed form.

# Outline

- 1 Discriminative Models / Logistic Regression
- 2 Bayesian Logistic Regression**
- 3 Generative Models
- 4 Assessing prediction in classification

# Laplace approximation / Bayesian CLT

- Given data  $y = (y_1, \dots, y_n)$  denote the **likelihood**  $f(y|\theta)$ .
- The prior  $\pi(\theta)$  could be improper but we assume that the posterior is proper and that its **mode exists**.
- Let  $\pi^*(\theta|x) = f(x|\theta)\pi(\theta)$  and denote the posterior mode  $\theta_M$ , which is (under regularity conditions) a solution of

$$\nabla_{\theta} \log \pi^*(\theta_M|x) = 0, \text{ for all } i = 1, \dots, p.$$

Also, let  $H(\theta)$  be the **Hessian** matrix.

- Then as  $n \rightarrow \infty$

$$\pi(\theta|x) \rightarrow N\left(\theta_M, H^{-1}(\theta_M)\right)$$

**Proof:** Similar to that of the **asymptotic** distribution of MLEs.

# Bayesian CLT - Example 1: Binomial

- Let  $y$  be an observation from a Binomial( $n, \theta$ ) and  $\pi(\theta) \propto 1$ .
- The mode can be found as  $\theta_M = y/n$ .
- The Hessian is equal to

$$H(\theta) = \frac{y}{\theta^2} + \frac{n-y}{(1-\theta)^2}$$

- Then as  $n \rightarrow \infty$

$$\pi(\theta|y) \rightarrow N\left(\frac{y}{n}, \frac{\frac{y}{n}(1 - \frac{y}{n})}{n}\right)$$

# Bayesian Logistic Regression - Laplace approximation

- Let's return to the **logistic regression** model. Assign the Normal prior on  $\beta$  with mean  $\beta_0$  and covariance  $\Sigma_0$ .

- We now need to **maximise**

$$\log(\pi(\beta|y, X)) = \log f(y|X, \beta) - \frac{1}{2}(\beta - \beta_0)^T \Sigma_0^{-1}(\beta - \beta_0)$$

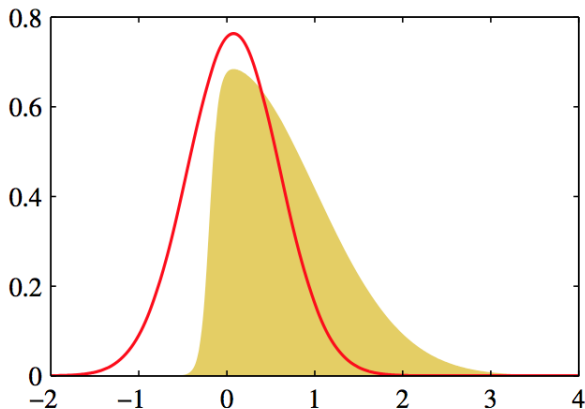
- The **Laplace approximation** of the posterior then becomes

$$N \left[ \beta_M, \left( \Sigma_0^{-1} + H(\beta_M) \right)^{-1} \right]$$

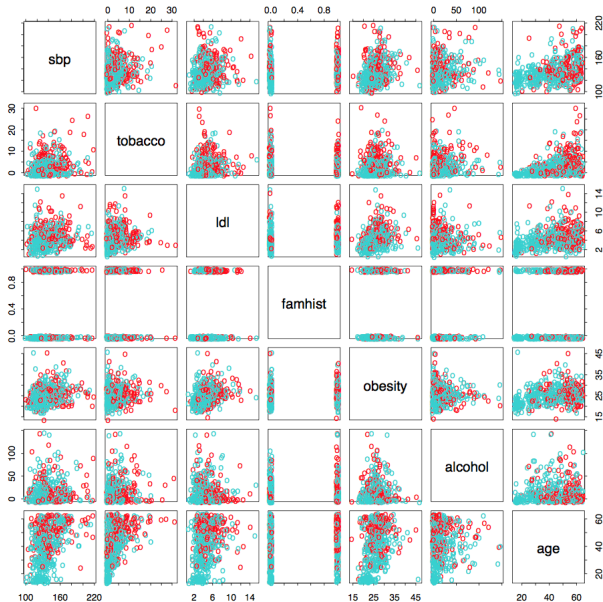
- This approximation will work well for **sufficiently large**  $n$ . We will say better approximations in the following weeks.

# Laplace Approximation

Below is a graphical illustration of the Laplace approximation. See also Exercise 2.



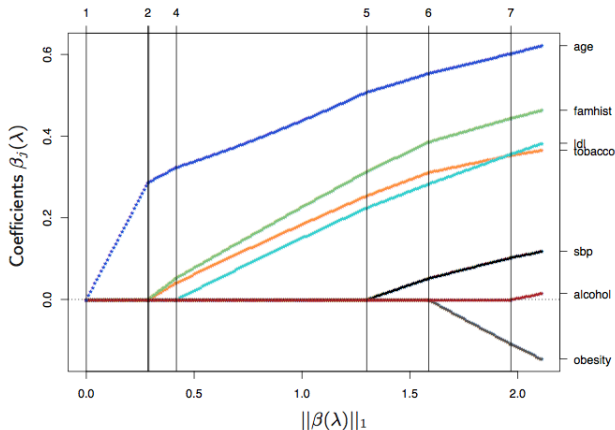
# Example: South African Heart Disease Data





# Laplace Approximation

As with linear regression **Lasso** and **Ridge** are special cases of the Bayesian approach by setting the corresponding **priors**. The Lasso results are shown below:



## Model Choice

The Laplace approximation can also be viewed as **Taylor expansion** around  $\beta_M$

$$\pi(\beta|y) \approx f(y|\beta_M)\pi(\beta_M) \exp\left(-\frac{1}{2}(\beta - \beta_M)^T H(\beta_M)(\beta - \beta_M)\right)$$

The **model evidence / marginal likelihood**  $\pi(y)$  is the normalising constant of  $\pi(\beta|y)$  so it may be approximated by

$$\begin{aligned}\pi(y) &\approx \int f(y|\beta_M)\pi(\beta_M) \exp\left(-\frac{1}{2}(\beta - \beta_M)^T H(\beta_M)(\beta - \beta_M)\right) d\beta \\ &= f(y|\beta_M)\pi(\beta_M)(2\pi)^{p/2} |H(\beta_M)|^{-1/2}\end{aligned}$$

Another approximation (not using priors) is offered by the **Bayesian Information Criterion (BIC)**

$$\log \pi(y) \approx \log f(y|\beta_M) - \frac{1}{2}p \log n$$

# Bayesian Logistic Regression - predictive distribution

Given a new set of covariate  $X_n$ , we can forecast  $y_n$  via the **predictive distribution**. Based on the Laplace approximation we can write

$$\pi(y_n|X_n, y, X) \approx \int \text{Bernoulli}(\sigma(X\beta)) N \left[ \beta_M, \left( \Sigma_0^{-1} + H_\beta \right)^{-1} \right] d\beta$$

The integral above **cannot be computed** analytically but we can sample from  $\pi(y_n|X_n, y, X)$  by

- 1 Draw  $N$  **Monte Carlo samples**  $\beta^i, i = 1, \dots, N$  from  $N \left[ \beta_M, \left( \Sigma_0^{-1} + H_\beta \right)^{-1} \right]$
- 2 obtain **predictive probabilities** by averaging the  $E[\sigma(X\beta^i)]$ 's

# Outline

- 1 Discriminative Models / Logistic Regression
- 2 Bayesian Logistic Regression
- 3 Generative Models**
- 4 Assessing prediction in classification

# Generative Models

The key difference with logistic regression is that we now specify a distribution for **both  $X$  and  $y$**  in the following way

$$\pi_{\theta}(y = c_k, X) = \pi_{\theta_y}(y = c_k)\pi_{\theta_x}(X|y = c_k), \quad k = 1, \dots, K.$$

The equation above is useful for **training** purposes. If  $\theta = (\theta_x, \theta_y)$  is not treated in a Bayesian manner, we use the MLE  $\hat{\theta}$ .

For **prediction** purposes we can use Bayes theorem to **forecast**  $y_n$  for a new point  $X_n$

$$\pi_{\hat{\theta}}(y_n = c_k|X_n) = \frac{\pi_{\hat{\theta}_x}(X_n|y = c_k)\pi_{\hat{\theta}_y}(y = c_k)}{\sum_{k=1}^K \pi_{\hat{\theta}_x}(X_n|y = c_k)\pi_{\hat{\theta}_y}(y = c_k)}, \quad k = 1, \dots, K.$$

## Softmax and discriminant function

Setting  $a_k(X) = \log \left[ \pi_{\hat{\theta}_x}(X|c_k) \pi_{\hat{\theta}_y}(c_k) \right]$ , we get the **softmax** function

$$\pi(c_k|X) = \frac{\exp(a_k(X))}{\sum_{k=1}^K \exp(a_k(X))}$$

which is  $\approx 1$  when  $a_k \gg a_j$  for all  $k \neq j$ .

In the case of two classes we the **logistic sigmoid**

$$\pi(c_k|X) = \frac{1}{1 + \exp(-d(X))} = \sigma(d(X)),$$

for the **discriminant function**  $d(X)$ ,

$$d(X) = \log \left( \frac{\pi_{\hat{\theta}_x}(X|c_1) \pi_{\hat{\theta}_y}(c_1)}{\pi_{\hat{\theta}_x}(X|c_0) \pi_{\hat{\theta}_y}(c_0)} \right)$$

## Example: Linear discriminant analysis

Assume **two classes**  $y = 0$  or  $y = 1$  and that the inputs  $X$  are  $N(\mu_0, \Sigma)$  if  $y = 0$  and  $N(\mu_1, \Sigma)$  if  $y = 1$ . Also  $P(y = 1) = \pi$ , so  $P(y = 0) = 1 - \pi$ .

The **likelihood** for  $\theta = (\pi, \mu_1, \mu_2, \Sigma)$  based  $(y_i, X_i)_{i=1}^n$  can be written as

$$\pi(X, y | \theta) = \prod_{i=1}^n [\pi N(\mu_1, \Sigma)]^{y_i} [(1 - \pi) N(\mu_0, \Sigma)]^{1-y_i}$$

Standard techniques yield the following **MLEs** of  $\theta$  (see exercise 1):

$$\hat{\pi} = \frac{n_1}{n}, \quad \text{where } n_1 \text{ is the number of points in class 1}$$

$$\hat{\mu}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} y_i X_{1i} \quad \hat{\mu}_0 = \frac{1}{n_0} \sum_{i=1}^{n_0} (1 - y_i) X_{0i}$$

$$\hat{\Sigma} = \frac{1}{n} \left( \sum_{i=1}^{n_1} (X_{1i} - \mu_1)(X_{1i} - \mu_1)^T + \sum_{i=1}^{n_2} ([-\mu_0](X_{0i} - \mu_0))^T \right)$$

# Notes on Linear discriminant analysis

- The parameter  $\mu_k$  refers to the profile of a **typical individual** in class  $k$
- We can write  $\pi(y = 1|X) = \sigma(\beta X + C)$  for some  $\beta = \Sigma^{-1}(\mu_1 - \mu_0)$  and a constant  $C$ , hence the discriminant function is **linear**.
- In case of **different**  $\Sigma$ 's for each class we get a quadratic discriminant function **quadratic discriminant analysis**
- **Fully Bayesian** inference on  $\theta$  can be made by assigning appropriate priors and deriving the posterior. Not pursued here.
- In case of  $p$  discrete  $X$ 's (binary features) we have  **$2^p$  cases**. Usually independent  $X$ 's are assumed to reduce the number of cases. This is called **naive Bayes**.



# Outline

- 1 Discriminative Models / Logistic Regression
- 2 Bayesian Logistic Regression
- 3 Generative Models
- 4 Assessing prediction in classification

# Sensitivity, specificity and misclassification rate

To **classify a new** individual with  $X_n$ , we can use  $\pi(y = 1|X_n)$ .

Two types of error: **False positives** and **False negatives**. If equally important the optimal prediction rule classifies  $y = 1$  if  $\pi(y = 1|X_n) > 0.5$ . Then check the **misclassification/accuracy rate**.

Different thresholds can also be used. Below are the in-sample **confusion matrices** for LDA in the 'Default' dataset with threshold 0.5

		<i>True default status</i>		
		No	Yes	Total
<i>Predicted default status</i>	No	9,644	252	9,896
	Yes	23	81	104
Total		9,667	333	10,000

**Sensitivity:**  $81/333 = 0.24$ , **Specificity:**  $9644/9667 = 0.99$ .

## Sensitivity, specificity and misclassification rate (cont'd)

But if we set a **lower** threshold of 0.2 we get

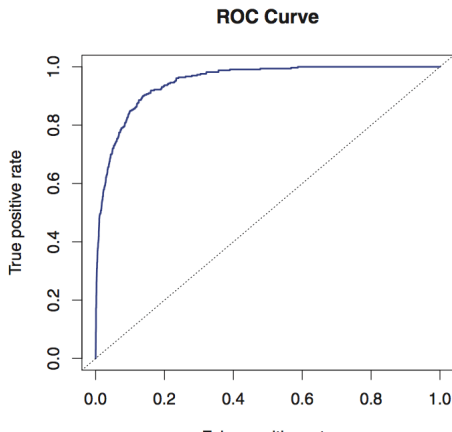
		<i>True default status</i>		
		No	Yes	Total
<i>Predicted default status</i>	No	9,432	138	9,570
	Yes	235	195	430
	Total	9,667	333	10,000

**Sensitivity:**  $195/333 = 0.59$ , **Specificity:**  $9432/9667 = 0.97$ .

So **which threshold** should we use when comparing models?

# ROC curves

For an overall measure we can look at the **area under the ROC curve** (sensitivity vs 1-specificity). In this case it is 0.95 which is quite good (0.5 corresponds to random guessing).



# Evaluating probabilistic forecasts - Scoring rules

- Let's say it didn't rain today. One model predicted rain with 0.99 probability and another with 0.51. Which of the two is **better**?
- **Scoring rules** are often used to evaluate probabilistic forecasts.
- Imagine a model that captures the probabilities of nature perfectly. If under a scoring rule, this model attains the optimal performance then the scoring rule is called **proper**. If this can only happen by this model, the rule is called **strictly proper**.
- Misclassification error is **not even a scoring rule** as it doesn't take into account these probabilities. Area under the ROC is **approximately** proper.

## Strictly proper scoring rules

- The **log score**,  $LS = -\log f(y|\pi)$  with  $f(\cdot)$  denoting the likelihood/density, is an example of a strictly proper scoring rule.
- If it didn't rain today, for model that predicted rain with  $\pi = 0.51$  it takes the value

$$LS = -\log \left[ 0.51^0 (1 - 0.51)^1 \right] = -\log(0.49) = 0.71$$

- For the model that predicted rain with  $p = 0.99$ , it takes the value

$$LS = -\log \left[ 0.99^0 (1 - 0.99)^1 \right] = -\log(0.01) = 4.61.$$

- **Smaller values** of LS are better so the model with  $\pi = 0.51$  scores better.

# Today's lecture - Reading

Bishop: 4.2 to 4.5.

Murphy: 4.2.1 to 4.2.4 8.1 8.2 8.3.1 8.3.3 8.3.7 and 8.4.1 to 8.4.4.