582631 — 5 credits

Introduction to Machine Learning

Lecturer: Teemu Roos
Assistant: Ville Hyvönen

Department of Computer Science
University of Helsinki

(based in part on material by Patrik Hoyer and Jyrki Kivinen)

November 1st–December 16th 2016

Classification: Probabilistic Methods

Logistic regression

- Logistic regression models are linear models for probabilistic binary classification (so, not really regression where response is continuous)
- Given input (vector) $\mathbf{x}$, the output is a probability that $Y = 1$
- However, instead of using a linear model directly as in

$$\Pr(Y = 1 \mid \mathbf{x}) = \boldsymbol{\beta} \cdot \mathbf{x}$$

  we let

$$\log \frac{\Pr(Y = 1 \mid \mathbf{x})}{\Pr(Y = 0 \mid \mathbf{x})} = \boldsymbol{\beta} \cdot \mathbf{x}$$

- This amounts to the same as

$$\Pr(Y = 1 \mid \mathbf{x}) = \frac{\exp(\boldsymbol{\beta} \cdot \mathbf{x})}{1 + \exp(\boldsymbol{\beta} \cdot \mathbf{x})} = \frac{1}{\exp(-\boldsymbol{\beta} \cdot \mathbf{x}) + 1}$$

Logistic regression (2)

- For convenience, we use here class labels 0 and 1
- Given probabilistic prediction $\hat{p}(y \mid \mathbf{x})$, and assuming instance $\mathbf{x}_i$ has already been observed, the **conditional likelihood** for a sample point $(\mathbf{x}_i, y_i)$ is

$$\begin{aligned} \hat{p}(Y = 1 \mid \mathbf{x}_i) \quad & \text{if} \quad y_i = 1 \\ 1 - \hat{p}(Y = 1 \mid \mathbf{x}_i) \quad & \text{if} \quad y_i = 0 \end{aligned}$$

which we write as

$$\hat{p}(Y = 1 \mid \mathbf{x}_i)^{y_i}(1 - \hat{p}(Y = 1 \mid \mathbf{x}_i))^{1-y_i}$$

Logistic regression (3)

- Conditional likelihood of sequence of independent samples (x_i, y_i), $i = 1, \ldots, n$ is then $\prod_{i=1}^{n} \hat{p}(Y = 1 \mid \mathbf{x}_i)^{y_i}(1 - \hat{p}(Y = 1 \mid \mathbf{x}_i))^{1-y_i}$
  - we say 'conditional' to emphasise that we take $\mathbf{x}_i$ as given and only model probability of labels y_i

- To maximise conditional likelihood, we can equivalently maximise conditional log-likelihood

$$\begin{aligned} \mathrm{LCL}(\beta)) &= \ln \prod_{i=1}^{n} \hat{p}(Y = 1 \mid \mathbf{x}_i)^{y_i}(1 - \hat{p}(Y = 1 \mid \mathbf{x}_i)^{1-y_i}) \\ &= \sum_{i=1}^{n} (y_i \ln \hat{p}(Y = 1 \mid \mathbf{x}_i) + (1 - y_i) \ln(1 - \hat{p}(Y = 1 \mid \mathbf{x}_i \end{aligned}$$

- Note that this is the same as **log-loss**!

Logistic regression (4)

- Maximizing the likelihood (or minimizing log-loss) isn't as straightforward as in the case of linear regression
- Nevertheless, the problem is convex which means that gradient-based techniques exist to find the optimum
- Standard techniques in R, Python, Matlab, ...
- Often used with regularisation, as in linear regression
  - "ridge": $\arg\max(\mathrm{LCL}(\boldsymbol{\beta}) - \lambda \|\boldsymbol{\beta}\|_2^2)$
  - "lasso": $\arg\max(\mathrm{LCL}(\boldsymbol{\beta}) - \lambda \|\boldsymbol{\beta}\|_1)$
- In particular, if data is linearly separable, non-regularised solution tends to infinity

Generative vs discriminative learning

- Logistic regression was an example of a **discriminative** and **probabilistic** classifier that directly models the class distribution $P(y \mid \mathbf{x})$
- Another probabilistic way to approach the problem is to use **generative** learning that builds a model for the whole joint distribution $P(\mathbf{x}, y)$ — often using the decomposition $P(y)P(\mathbf{x} \mid y)$
- Both approaches have their pros and cons:
  - Discriminative learning: only solve the task that you need to solve; may provide better accuracy since focuses on the specific learning task; optimization tends to be harder
  - Generative learning: often more natural to build models for $P(\mathbf{x} \mid y)$ than for $P(y \mid \mathbf{x})$; handles missing data more naturally; optimization often easier

Generative vs discriminative learning (2)

- Estimating the *class prior* $P(y)$ is usually simple

- For example, in binary classification — this time with $Y \in \{-1, +1\}$ — we can usually just count the number of positive examples *Pos* and negative examples *Neg* and set

$$P(Y = +1) = \frac{Pos}{Pos + Neg} \quad \text{and} \quad P(Y = -1) = \frac{Neg}{Pos + Neg}$$

- Since $P(\mathbf{x}, y) = P(\mathbf{x} \mid y)P(y)$, what remains is estimating $P(\mathbf{x} \mid y)$. In binary classification, we
  - use the positive examples to build a model for $P(\mathbf{x} \mid Y = +1)$
  - use the negative examples to build a model for $P(\mathbf{x} \mid Y = -1)$

- To classify a new data point $\mathbf{x}$, we use the Bayes formula

$$P(y \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid y)P(y)}{P(\mathbf{x})} = \frac{P(\mathbf{x} \mid y)P(y)}{\sum_{y'} P(\mathbf{x} \mid y')P(y')}$$

Generative vs discriminative learning (3)

Examples of discriminative classifiers:

- logistic regression
- k-NN
- decision trees
- SVM
- multilayer perceptron (MLP)

Examples of generative classifiers:

- naive Bayes (NB)
- linear discriminant analysis (LDA)
- quadratic discriminant analysis (QDA)

We will study all of the above except MLP.

Normal distribution

- For probabilistic models for real-valued features $x_i \in \mathbb{R}$, one basic ingredient is the *normal* or *Gaussian* distribution
- Recall that for a single real-valued random variable, the normal distribution has two parameters μ and σ^2, and density

$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

  - If X has this distribution, then $\mathrm{E}[X] = \mu$ and $\mathrm{Var}[X] = \sigma^2$
- For multivariate case $\mathbf{x} \in \mathbb{R}^p$, we shall first consider the case where individual component x_i has normal distribution with parameters μ_i and σ_i^2 and the components are independent:

$$p(\mathbf{x}) = \mathcal{N}(x_1 \mid \mu_1, \sigma_1^2) \ldots \mathcal{N}(x_p \mid \mu_p, \sigma_d^2)$$

Normal distribution (2)

- We get

$$
\begin{aligned}
p(\mathbf{x}) &= \mathcal{N}(x_1 \mid \mu_1, \sigma_1^2) \ldots \mathcal{N}(x_p \mid \mu_p, \sigma_p^2) \\
&= \prod_{j=1}^{p} \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{(x_j - \mu_j)^2}{2\sigma_j^2}\right) \\
&= \frac{1}{(2\pi)^{p/2}\sigma_1 \ldots \sigma_p} \exp\left(-\frac{1}{2}\sum_{j=1}^{p} \frac{(x_j - \mu_j)^2}{\sigma_j^2}\right) \\
&= \frac{1}{(2\pi)^{p/2}\,|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)
\end{aligned}
$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p) \in \mathbb{R}^p$ and $\Sigma \in \mathbb{R}^{p \times p}$ is a diagonal matrix with $\sigma_1^2, \ldots, \sigma_p^2$ on the diagonal and $|\Sigma|$ is determinant of Σ

Normal distribution (3)

- More generally, let $\boldsymbol{\mu} \in \mathbb{R}^p$, and let $\Sigma \in \mathbb{R}^{p \times p}$ be
  - symmetric: $\Sigma^{\mathrm{T}} = \Sigma$
  - positive definite: $\mathbf{x}^{\mathrm{T}} \Sigma \mathbf{x} > 0$ for all $\mathbf{x} \in \mathbb{R} - \{ 0 \}$

- We then define p-dimensional Gaussian density with parameter $\boldsymbol{\mu}$ and Σ as

$$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{p/2} \left|\Sigma\right|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right)$$

- If Σ is diagonal, we get the special case where x_j are independent

Normal distribution (4)

- To understand the multivariate normal distribution, consider a surface of constant density:

$$S = \{\, \mathbf{x} \in \mathbb{R}^p \mid \mathcal{N}(\mathbf{x} \mid \mu, \Sigma) = a \,\}$$

for some a

- By definition of $\mathcal{N}$, this can be written as

$$S = \left\{\, \mathbf{x} \in \mathbb{R}^p \mid (\mathbf{x} - \mu)^{\mathrm{T}} \Sigma^{-1} (\mathbf{x} - \mu) = b \,\right\}$$

for some b

- Because Σ is symmetric and positive definite, so is Σ^{-1}, and this set is an ellipsoid with centre μ

Normal distribution (5)

- More specifically, since Σ is symmetric and positive definite, it has an Eigenvalue decomposition

$$\Sigma = U\Lambda U^{\mathrm{T}}$$

where $\Lambda \in \mathbb{R}^{p\times p}$ is diagonal and $U \in \mathbb{R}^{\mathrm{T}}$ is orthogonal $(U^{\mathrm{T}} = U^{-1})$, and further

$$\Sigma^{-1} = U\Lambda^{-1}U^{\mathrm{T}}$$

- We then know from analytic geometry that for the ellipsoid

$$S = \left\{ \mathbf{x} \in \mathbb{R}^p \mid (\mathbf{x} - \mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x} - \mu) = b \right\}$$

  - the directions of the axes are given by the column vectors of U (Eigenvectors of Σ)
  - the squared lengths of the axes are given by the elements of Λ (Eigenvalues of Σ)

Normal distribution (6)

- Let $\mathbf{X} = (X_1, \ldots, X_p)$ have normal distribution with parameters $\boldsymbol{\mu}$ and Σ
- Then $\mathrm{E}[\mathbf{X}] = \boldsymbol{\mu}$ and $\mathrm{E}[(X_r - \mu_r)(X_s - \mu_s)] = \Sigma_{rs}$
- Hence, we call the parameter $\boldsymbol{\mu}$ the mean and Σ the covariance matrix

Normal distribution (7)

- Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$, where $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,p})$, be n independent samples from a p-dimensional normal distribution with unknown mean $\boldsymbol{\mu}$ and covariance Σ

- The maximum likelihood estimates

$$(\hat{\boldsymbol{\mu}}, \hat{\Sigma}) = \arg \max_{\boldsymbol{\mu}, \Sigma} \prod_{i=1}^{n} \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}, \Sigma)$$

are given by

$$\begin{aligned} \hat{\mu}_r &= \frac{1}{n} \sum_{i=1}^{n} x_{i,r} \\ \hat{\Sigma}_{rs} &= \frac{1}{n} \sum_{i=1}^{n} (x_{i,r} - \hat{\mu}_r)(x_{i,s} - \hat{\mu}_s) \end{aligned}$$

Gaussians in classification

- LDA and QDA are obtained by modeling positive and negative examples both with their own Gaussian:

$$
\begin{aligned}
p(\mathbf{x} \mid Y = +1) &= \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_+, \Sigma_+) \\
p(\mathbf{x} \mid Y = -1) &= \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_-, \Sigma_-))
\end{aligned}
$$

  where $\boldsymbol{\mu}_\pm$ and $\Sigma_\pm$ are obtained for example as maximum likelihood estimates

- Decision boundary is given by

$$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_+, \Sigma_+) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_-, \Sigma_-)$$

  or equivalently

$$\ln \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_+, \Sigma_+) = \ln \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_-, \Sigma_-)$$

Gaussians in classification (2)

- By substituting the formula for $\mathcal{N}$ into

$$\ln \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_+, \Sigma_+) = \ln \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_-, \Sigma_-)$$

  and simplifying we get

$$(\mathbf{x}-\boldsymbol{\mu}_+)^T \Sigma_+^{-1}(\mathbf{x}-\boldsymbol{\mu}_+) - (\mathbf{x}-\boldsymbol{\mu}_-)^T \Sigma_-^{-1}(\mathbf{x}-\boldsymbol{\mu}_-) + \ln \frac{|\Sigma_-|}{|\Sigma_+|} = 0$$

- If $\Sigma_+ = \Sigma_-$ this is a linear equation, so the decision boundary is a hyperplane: **LDA**
- In general case this is a quadratic surface: **QDA**
- In QDA, decision regions may be non-connected