GENERALISATION BOUNDS FOR GENERATIVE/DISCRIMINATIVE LEARNING I

DEEP LEARNING (SS2020)
1. COMPUTER LAB (10P)

1. INTRODUCTION

This is the first of two home works aiming at an experimental comparison of generalisation bounds for generative and discriminative learning. We will consider a simple Gaussian classification problem and two learning approaches – generative learning by maximum likelihood estimate and discriminative learning by logistic regression.

2. MODEL & NOTATIONS

We consider a Gaussian classification problem with features $\boldsymbol{X} \in \mathbb{R}^d$ and two classes $Y \in \{0, 1\}$. The joint distribution over features and classes is defined by the prior class probabilities and the distributions for the features conditioned on the classes, which are assumed as multivariate normal distributions.

$$\begin{gathered} \boldsymbol{X} \mid Y \sim \mathcal{N}(\boldsymbol{\mu}_Y, \boldsymbol{C}_Y) \\ \mathbb{P}(Y=0) = \pi_0 \, , \mathbb{P}(Y=1) = \pi_1 . \end{gathered} \tag{1}$$

Here, we will assume that the covariance matrices of the two distributions are identical, i.e. $\boldsymbol{C}_0 = \boldsymbol{C_1} = \boldsymbol{C}$. A linear classifier with parameters $\boldsymbol{w} \in \mathbb{R}^d$, $b \in \mathbb{R}$ is defined as $h(x) = \mathrm{H}(\boldsymbol{w}^T \boldsymbol{x} + b)$ where H denotes the Heaviside function. Using standard 0/1-loss, its risk is given by

$$R(h) = 1 - \pi_1 \Phi\Big(\frac{\boldsymbol{w}^T \boldsymbol{\mu_1} + b}{\sqrt{\boldsymbol{w}^T \boldsymbol{C} \boldsymbol{w}}}\Big) - \pi_0 \Phi\Big(-\frac{\boldsymbol{w}^T \boldsymbol{\mu_0} + b}{\sqrt{\boldsymbol{w}^T \boldsymbol{C} \boldsymbol{w}}}\Big), \tag{2}$$

where Φ denotes the cdf of the standard univariate normal distribution. The optimal classifier is the Bayes classifier $h^*(x) = \mathrm{H}\big(\frac{\pi_1 \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{C})}{\pi_0 \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{C})} - 1\big)$, which is a linear classifier

$$h^*(\boldsymbol{x}) = \mathrm{H}\Big[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \boldsymbol{C}^{-1} \boldsymbol{x} + \frac{1}{2}\big(\boldsymbol{\mu}_0^T \boldsymbol{C}^{-1} \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1^T \boldsymbol{C}^{-1} \boldsymbol{\mu}_1\big) + \log \frac{\pi_1}{\pi_0}\Big]. \tag{3}$$

A classifier is trained using i.i.d training sets $\mathcal{T}^m = \{(\boldsymbol{x}^j, y^j) \mid j = 1, \dots, m\}$

3. GENERATIVE LEARNING

Generative learning is straightforward here. Given training data $\mathcal{T}^m$, we estimate the model parameters $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$, the covariance matrix $\boldsymbol{C}$ and the prior class probabilities $\pi_{0/1}$. Given the estimate, the predictor h_m is obtained from (3).

4. Discriminative learning

We will use the logistic regression approach. Using the ± 1 encoding of the two classes obtained by $y \to 2y - 1$, it can be shown that the conditional class probabilities of model (1) have the form

$$p(y \mid \boldsymbol{x}) = \frac{e^{y(\boldsymbol{w}^T\boldsymbol{x}+b)}}{2\cosh(\boldsymbol{w}^T\boldsymbol{x}+b)} = \mathrm{S}\big(y(\boldsymbol{w}^T\boldsymbol{x}+b)\big), \tag{4}$$

where S is the sigmoid function. The parameters $(\boldsymbol{w}, b)$ can be computed form the model parameters of (1).

The conditional log-likelihood of the training data $\mathcal{T}^m$ is

$$\frac{1}{m}\sum_{j=1}^{m} \log \mathrm{S}\big(y^j(\boldsymbol{w}^T\boldsymbol{x}^j+b)\big). \tag{5}$$

We will simplify this expression in two steps. First, by extending $\boldsymbol{x}$ and $\boldsymbol{w}$ by an additional dimension, and second, redefining $\boldsymbol{x}^j \to -y^j\boldsymbol{x}^j$ and, finally, switching to minimisation of the negative conditional log-likelihood, we get the task

$$L(\boldsymbol{w}) = \frac{1}{m}\sum_{j=1}^{m} \log(1+e^{\boldsymbol{w}^T\boldsymbol{x}^j}) \to \min_{\boldsymbol{w}}. \tag{6}$$

Its objective function is convex and differentiable in $\boldsymbol{w}$ and can be minimised by gradient descent.

5. Assignments

Assignment 1. (1p)
Give formulas for maximum likelihood estimates of the model parameters $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$, $\boldsymbol{C}$ and $\pi_{0/1}$ from training data $\mathcal{T}^m = \{(\boldsymbol{x}^j, y^j) \mid j = 1, \dots, m\}$.

Assignment 2. (2p)
Deduce the risk formula (2) for a linear classifier parametrised by $(\boldsymbol{w}, b)$. Hints:

(1) Start from the following fact. If the random vector $\boldsymbol{X}$ has multivariate normal distribution $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{C})$, then the scalar random variable $Z = \boldsymbol{w}^T\boldsymbol{X} + b$ has normal distribution. Find its mean and variance.

(2) Use the fact that the cdf of a normal distribution with mean μ and variance σ^2 is given by $\Phi(\frac{x-\mu}{\sigma})$, where Φ denotes the cdf of the standard normal distribution with zero mean and unit variance.

Assignment 3. (2p)
Prove that the posterior class probabilities of model (1) have the form (4).

Assignment 4. (3p)
Implement the gradient descent optimisation for the logistic regression (6). Notice that the objective function has Lipschitz continuous gradient

$$\|\nabla L(\boldsymbol{w}) - \nabla L(\boldsymbol{w}')\| < G\|\boldsymbol{w} - \boldsymbol{w}'\|,$$

where $G = \frac{1}{m}\sum_j \|\boldsymbol{x}^j\|^2$. It is therefore “safe” to use a constant step width $\alpha < \frac{1}{G}$ and the stopping criterion $\|\nabla L(\boldsymbol{w})\| < 0.5\alpha\epsilon$ with $\epsilon = \|\boldsymbol{w} - \boldsymbol{w}^*\|$ representing the desired distance from the optimiser $\boldsymbol{w}^*$. Propose a reasonable initialisation of $\boldsymbol{w}$.

Assignment 5. (2p)

Fix a model (1). You may use the four-dimensional model provided at the course web page. The npz-archive contains all model parameters. The corresponding keys of the dictionary are `['clps', 'mues', 'cov', 'dim']`. Report the risk R_B of the Bayes optimal predictor.

Estimate the generalisation errors for the two learning approaches for training set sizes $m = 50, 100, 500, 1000, 10000$. For each m repeat the following experiment 500 times. Generate an i.i.d. training set from the true model. Learn the predictors h_m by the two considered approaches. Compute the average excess risk $R(h_m) - R_B$ and its standard deviation over the repetitions.

Report (graphically!) the dependence of the average excess risk and its standard deviation on the training set size. Compare it for the two methods.