IST Austria: Statistical Machine Learning 2015/16
Christoph Lampert <chl@ist.ac.at>
TA: Alex Zimin <azimin@ist.ac.at>
Exercise Sheet 2/5

1 Bayes Classifier

In the lecture we saw that the Bayes classifier is

$$c^*(x) := \operatorname{argmax}_{y\in\mathcal{Y}} p(y|x). \tag{1}$$

a) Which of these decision functions is equivalent to c^*?

- $c_1(x) := \operatorname{argmax}_y p(x)$
- $c_2(x) := \operatorname{argmax}_y p(y)$
- $c_3(x) := \operatorname{argmax}_y p(x, y)$
- $c_4(x) := \operatorname{argmax}_y p(x|y)$

For $\mathcal{Y} = \{-1, +1\}$, we can express the Bayes classifier as $c^*(x) = \operatorname{sign}[\log \frac{p(+1|x)}{p(-1|x)}]$
b) Which of the following expressions are equivalent to c^*?

- $c_5(x) := \operatorname{sign}[\frac{\log p(x,+1)}{\log p(x,-1)}]$
- $c_6(x) := \operatorname{sign}[\log p(+1|x) + \log p(-1|x)]$
- $c_7(x) := \operatorname{sign}[\log p(+1|x) - \log p(-1|x)]$
- $c_8(x) := \operatorname{sign}[\log p(x, +1) - \log p(x, -1)]$
- $c_9(x) := \operatorname{sign}[p(+1|x) - p(-1|x)]$
- $c_{10}(x) := \operatorname{sign}[\frac{p(x,+1)}{p(x,-1)} - 1]$
- $c_{11}(x) := \operatorname{sign}[\frac{\log p(+1|x)}{\log p(-1|x)} - 1]$
- $c_{12}(x) := \operatorname{sign}[\log \frac{p(x|+1)}{p(x|-1)} + \log \frac{p(+1)}{p(-1)}]$

2 Gaussian Discriminant Analysis

Gaussian Discriminant Analysis (GDA) is an easy-to-compute method for generative probabilistic classification. For a training set $\mathcal{D} = \{(x^1, y^1), \dots, (x^n, y^n)\}$ set

$$\mu := \frac{1}{n}\sum_{i=1}^{n} x^i, \qquad \Sigma := \frac{1}{n}\sum_{i=1}^{n}(x^i - \mu)(x^i - \mu)^\top, \qquad \mu_y := \frac{1}{|\{i : y^i = y\}|}\sum_{\{i:y^i=y\}} x^i, \quad \text{for } y \in \mathcal{Y}, \tag{2}$$

and define

$$p(x|y) = \frac{1}{\sqrt{2\pi \det \Sigma}} \exp(-\frac{1}{2}(x - \mu_y)^\top \Sigma^{-1} (x - \mu_y)) \tag{3}$$

a) Show for binary classification tasks: GDA leads to a linear decision rule, regardless of what $p(y)$ is.
b) GDA is popular when there are many classes but only few examples for each class. Can you imagine why?

3 Robustness of the Perceptron

Look at the dataset with the following three points:

$$\mathcal{D} = \{ (\begin{pmatrix}2\\1\end{pmatrix}, +1), (\begin{pmatrix}-1\\-2\end{pmatrix}, -1), (\begin{pmatrix}a\\b\end{pmatrix}, +1)\} \subset \mathbb{R}^2 \times \{\pm 1\}.$$

- For any $0 < \rho \leq 1$, find values for a and b such that the Perceptron algorithm converges to a *correct* classifier with *robustness* ρ.
- What's the maximal robustness you can achieve for any choice of a and b?

4 Perceptron Training as Convex Optimization

The following form of Perceptron training can be interpreted as optimizing a convex, but non-differentiable, objective function by stochastic gradient descent. What is the objective? What is the stepsize rule? Discuss advantages and shortcomings of this interpretation.

Algorithm 1 Randomized Perceptron Training

input linearly separable training set $\mathcal{D} = \{(x^1, y^1), \dots, (x^n, y^n)\} \subset \mathbb{R}^d \times \{\pm 1\}$

$w_1 \leftarrow 0$

for $t = 1, \dots, T$ **do**

  $(x, y) \leftarrow$ random example from $\mathcal{D}$

  if $y\langle w_t, x\rangle \leq 0$ **then**

    $w_{t+1} \leftarrow w_t + yx$

  else

    $w_{t+1} \leftarrow w_t$

  end if

end for

output w_{T+1}

5 Hard-Margin SVM Dual

Compute the dual optimization problem to the hard-margin SVM training problem:

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \quad \frac{1}{2}\|w\|^2 \quad \text{subject to} \quad y^i(\langle w, x^i\rangle + b) \geq 1, \qquad \text{for } i = 1, \dots, n.$$

6 Missing Proofs

- Let $f_1, \dots, f_K$ be differentiable at w_0 and let $f(w) = \max\{f_1(w), \dots, f_K(w)\}$. Let k be any index with $f_k(w_0) = f(w_0)$. Show that any v that is a subgradient of f_k at w_0 is also a subgradient of f at w_0.

- Let f be a convex function and denote by w^* a minimum of f. Let $w_{t+1} = w_t - \eta_t v$, where v is a subgradient of the f at w_t.

  Show: there exists a stepsize η_t such that $\|w_{t+1} - w^*\| < \|w_t - w^*\|$, except if w_t is a minimum already.

- In your above proof, w^* can be *any* minimum of f. Let w_1^* and w_2^* be two different minima, then w_t will converge towards both of them. Isn't this impossible?

  Note: this is not a trivial question: convex functions *can* have multiple global minima, e.g. $f(w) = 0$ has infinitely many.

- Let $g(\alpha) = \max_{\theta \in \Theta} f(\theta) + \sum_{i=1}^k \alpha_i g_i(\theta)$ be the dual function of an optimization problem.

  Show: g is always a convex function w.r.t. α, even if the original optimization problem was not convex.

7 Practical Experiments III

- Pick one more training methods from the previous sheet and implement it.
- In addition, implement a *linear support vector machine (SVM)* with training by the subgradient method.
- What error rates do both methods achieve on the datasets from the previous sheet?
- For the *wine* data, make a plot of the SVM's objective values and the Euclidean distance to the optimium (after you computed it in an earlier run) after each iteration.