

# 2413, Machine Learning, Tutorial 7

## Universität Bern

Simon Jenni (jenni@inf.unibe.ch)

### Mutual Information

1. Consider the following set of 2-dimensional points, sampled from two classes:

$x_1$	$x_2$	$y$
1	0	1
0	1	1
0	0	1
0	1	1
0	0	0
1	1	0
1	0	0
0	0	0

Find the feature  $x_i$  with the highest mutual information

$$MI(x_i, y) = \sum_{x_i \in \{0,1\}} \sum_{y \in \{0,1\}} p(x_i, y) \log \frac{p(x_i, y)}{p(x_i)p(y)}. \quad (1)$$

#### Solution

Since we don't know the true distributions of the variables, we use the empirical distributions:  $p(y = 1) = 0.5$ ,  $p(y = 0) = 0.5$ ,  $p(x_1 = 1) = 3/8$ ,  $p(x_1 = 0) = 5/8$ ,  $p(x_2 = 1) = 3/8$ ,  $p(x_2 = 0) = 5/8$ ,  $p(x_1 = 1, y = 1) = 1/8$ ,  $p(x_1 = 1, y = 0) = 2/8$ ,  $p(x_1 = 0, y = 1) = 3/8$ ,  $p(x_1 = 0, y = 0) = 2/8$ ,  $p(x_2 = 1, y = 1) = 2/8$ ,  $p(x_2 = 1, y = 0) = 1/8$ ,  $p(x_2 = 0, y = 1) = 2/8$ ,  $p(x_2 = 0, y = 0) = 3/8$ .

We then have:  $MI(x_1, y) = 1.29$  and  $MI(x_2, y) = 1.4$ . We can then conclude that  $x_2$  has the highest mutual information with the label  $y$ .

## Regularization

2. Consider the problem of linear regression with a Gaussian prior on the parameter vector  $\theta$  of the form  $p(\theta) = \mathcal{N}(0, \lambda I)$ , where  $I$  is the identity matrix. Derive the cost function for the **MAP** estimate  $\theta_{MAP}$ .

**Hint:** Remember the linear regression assumption  $y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)}$  where  $\epsilon^{(i)} \sim \mathcal{N}(0, \sigma)$ .

### Solution

Linear regression can be seen as a maximum likelihood estimation, where the error  $\epsilon^{(i)}$  is modeled according to a Gaussian Distribution:

$$p(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\epsilon^{(i)})^2}{2\sigma^2}\right). \quad (2)$$

that implies

$$p(y^{(i)}|x^{(i)}; \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right). \quad (3)$$

We then estimate the optimal  $\theta$  by maximizing eq.(3) over all the training samples

$$\theta_{ML} = \arg \max_{\theta} \prod_{i=1}^m p(y^{(i)}|x^{(i)}; \theta). \quad (4)$$

In the case of regularized linear regression  $\theta$  is a random variable, therefore we have that the optimal  $\theta$  is given by the *maximum a posteriori* estimate

$$\theta_{MAP} = \arg \max_{\theta} \prod_{i=1}^m p(y^{(i)}|x^{(i)}, \theta)p(\theta). \quad (5)$$

Since we assume that  $\theta_j$  are i.i.d. and follow a Gaussian distribution, we have

$$p(\theta) = \prod_{j=1}^n p(\theta_j) = \frac{1}{\sqrt{2\pi}\lambda} \prod_{j=1}^n \exp\left(-\frac{\theta_j^2}{2\lambda^2}\right). \quad (6)$$

since  $\arg \max_{\theta} \prod_{i=1}^m p(y^{(i)}|x^{(i)}, \theta)p(\theta) = \arg \max_{\theta} \log(\prod_{i=1}^m p(y^{(i)}|x^{(i)}, \theta)p(\theta))$ , we have

$$\theta_{MAP} = \arg \max_{\theta} \log\left(\prod_{i=1}^m p(y^{(i)}|x^{(i)}, \theta)p(\theta)\right) \quad (7)$$

$$= \arg \max_{\theta} m \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{1}{2\sigma^2} \sum_{i=1}^m (y^{(i)} - \theta^T x^{(i)})^2 + n \log\left(\frac{1}{\sqrt{2\pi}\lambda}\right) - \frac{1}{2\lambda^2} \sum_{j=1}^n \theta_j^2 \quad (8)$$

$$= \arg \min_{\theta} \frac{1}{2} \sum_{i=1}^m (y^{(i)} - \theta^T x^{(i)})^2 + \frac{\sigma^2}{2\lambda^2} \sum_{j=1}^n \theta_j^2. \quad (9)$$

3. What is the gradient of the cost derived in the previous question?

**Solution**

Eq.(9) can be written in the following notation

$$\ell(\theta; X, y, \hat{\lambda}) = \frac{1}{2} \|y - X\theta\|_2^2 + \frac{\hat{\lambda}}{2} \|\theta\|_2^2 \quad (10)$$

where  $X \in \mathcal{R}^{m \times n}$  is a matrix where the  $i$ th row is  $(x^{(i)})^T$ ,  $y \in \mathcal{R}^m$  is a vector where the  $i$ th element is  $y^{(i)}$  and  $\hat{\lambda} = \frac{\sigma^2}{\lambda^2}$ . The gradient to respect to  $\theta$  is then

$$\nabla_{\theta} \ell(\theta; X, y, \hat{\lambda}) = -X^T(y - X\theta) + \hat{\lambda}\theta. \quad (11)$$

If we equal the gradient to zero we obtain

$$\theta_{MAP} = (X^T X + \hat{\lambda} I)^{-1} X^T y \quad (12)$$