Machine Learning, Tutorial 8
Universität Bern

Abdelhak Lemkhenter (abdelhak.lemkhenter@inf.unibe.ch)

Bootsrap

1. Given a set $\mathcal{D}$, we use bootstrap to build a new set $\mathcal{Z}$ such that $|\mathcal{D}| = |\mathcal{Z}| = n$. Let v be a sample in $\mathcal{D}$:

   (a) What is the probability of v not being an element of $\mathcal{Z}$.

   (b) What is the probability of v not being an element of $\mathcal{Z}$ when $n \to +\infty$.
   Hint: $lim_{n\to+\infty}(1 - 1/n)^n = 1/e$.

Solution

(a) $\mathcal{Z}$ is built by drawing n samples from $\mathcal{D}$ with replacement. For every sample v' drawn from $\mathcal{D}$, $P(v' \neq v) = 1 - 1/n$. Therefore, $P(v \notin \mathcal{Z}) = (1 - 1/n)^n$.

(b) Using the hint, we obtain that the probability is $1/e$. The effective "size" of $\mathcal{Z}$ is $1 - 1/e \approx 63\%$.

Gradient boosting

2. Let $\mathcal{D} = \{(x_i, y_i)_{i=1}^n\}$ be a data set and $L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$. Show that the update step in the gradient boost algorithm is equivalent to equation 1.

$$\gamma_{k+1} = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} (e_i^k, G(x_i;\gamma)) \tag{1}$$
$$\text{where } e_i^k = G(x_i;\gamma_k) - y_i$$

Hint: The update step in the gradient boost algorithm is given by 2.

$$\gamma_{k+1} = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} (g_i^k, G(x_i;\gamma)) \tag{2}$$
$$\text{where } g_i^k = \frac{\partial L(y_i, G(x_i;\gamma_k))}{\partial G(x_i;\gamma_k))}$$

Solution
$\frac{\partial L(y,\hat{y})}{\partial \hat{y}} = \hat{y} - y$. Therefore, $g_i^k = e_i^k$. At each iteration, the new tree/model is learning to compensate for the error of the previous ones.

Mutual Information

3. 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

$$\mathrm{MI}\left(x_i, y\right) = \sum_{x_i \in \{0,1\}} \sum_{y \in \{0,1\}} p\left(x_i, y\right) \log \frac{p\left(x_i, y\right)}{p\left(x_i\right) p(y)}. \tag{3}$$

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

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

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

$$p(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\epsilon^{(i)})^2}{2\sigma^2}\right). \tag{4}$$

   that implies

$$p(y^{(i)}|x^{(i)};\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right). \tag{5}$$

   We then estimate the optimal θ by maximizing eq.(5) over all the training samples

$$\theta_{ML} = \arg\max_\theta \prod_{i=1}^m p(y^{(i)}|x^{(i)};\theta). \tag{6}$$

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

$$\theta_{MAP} = \arg\max_\theta \prod_{i=1}^m p(y^{(i)}|x^{(i)},\theta)p(\theta). \tag{7}$$

   Since we assume that θ_j are i.i.d. and follow a Gaussian distribution, we have

$$p(\theta) = \prod_{j=1}^n p(\theta_j) = \frac{1}{\sqrt{2\pi}\lambda} \prod_{j=1}^n \exp\left(-\frac{\theta_j^2}{2\lambda^2}\right). \tag{8}$$

   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

$$\begin{align}
\theta_{MAP} &= \arg\max_\theta \log(\prod_{i=1}^m p(y^{(i)}|x^{(i)},\theta)p(\theta)) \tag{9}\\
&= \arg\max_\theta m \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^m (y^{(i)} - \theta^T x^{(i)})^2 + n \log\left(\frac{1}{\sqrt{2\pi}\lambda}\right) - \frac{1}{2\lambda^2}\sum_{j=1}^n \theta_j^2 \tag{10}\\
&= \arg\min_\theta \frac{1}{2}\sum_{i=1}^m (y^{(i)} - \theta^T x^{(i)})^2 + \frac{\sigma^2}{2\lambda^2}\sum_{j=1}^n \theta_j^2. \tag{11}
\end{align}$$

5. What is the gradient of the cost derived in the previous question? **MAP** estimate θ_{MAP}. What advantage does θ_{MAP} have over θ_{ML}?

   Solution

   Eq.(11) can be written in the following notation

$$\ell(\theta; X, y, \hat{\lambda}) = \frac{1}{2}||y - X\theta||_2^2 + \frac{\hat{\lambda}}{2}||\theta||_2^2 \tag{12}$$

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

$$\nabla_\theta \ell(\theta; X, y, \hat{\lambda}) = -X^T(y - X\theta) + \hat{\lambda}\theta. \tag{13}$$

   If we equal the gradient to zero we obtain

$$\theta_{MAP} = (X^T X + \hat{\lambda} I)^{-1} X^T y \tag{14}$$

   θ_{MAP} is always defined even if X doesn't have a full rank since $(X^T X + \hat{\lambda} I)$ is positive definitive and thus invertible.