Course 4

November 21, 2013

1 Regularized Empirical Risk Minimization

Let us denote by $(y_i, x_i)_{i=1\cdots n}$ a set of training data where $y_i \in \{-1;1\}$ and $x_i \in \mathbb{R}^p$. We want to find $w \in \mathbb{R}^p$ such that $y_i \approx w^\top x_i$. To do so, we propose to minimize the sum of two terms w.r.t w. The first one is the empirical risk $R(w) = \frac{1}{n}\sum_{i=1}^n l\left(y_i; w^\top x_i\right)$ where l is called the loss function (in this course, it refers to a convex function). This term enforces a fit to data measurements The second term $\frac{\lambda}{2}\|w\|_2^2$ permits to introduce some regularity on w.

Proposition 1.1. *There is a "moral" equivalence between the following two problems :*

$$\widehat{w}(\lambda) = \arg\min_{w\in\mathbb{R}^p} R(w) + \frac{\lambda}{2}\|w\|_2^2 \tag{1}$$

$$\tilde{w}(T) = \arg\min_{w\in\mathbb{R}^p} R(w) \quad s.t. \quad \|w\|_2^2 \leq T \tag{2}$$

in a sense that for all λ, there exists $T > 0$ such that $\widehat{w}(\lambda) \subseteq \tilde{w}(T)$.

1.1 Ridge Regression

Ridge regression works for regularization problems and also for classifications problems. It is the specific case where $(\forall a \in \mathbb{R}), (\forall b \in \mathbb{R}),\ l(a,b) = \frac{1}{2}(a-b)^2$. The Ridge regression problem consists in finding the unique minimizer $\widehat{w}$ of the following quantity :

$$f(w) = \frac{1}{n}\sum_{i=1}^n \frac{1}{2}(y_i - w^\top x_i)^2 + \frac{\lambda}{2}\|w\|_2^2 \tag{3}$$

$$f(w) = \frac{1}{2n}\|y - Xw\|_2^2 + \frac{\lambda}{2}\|w\|_2^2 \tag{4}$$

where $y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n\times p}$. From first-order stationary condition :

$$\nabla f(w) = 0 \quad \Longleftrightarrow \quad -\frac{1}{n}X^\top[y - Xw] + \lambda w = 0 \tag{5}$$

$$(X^\top X + \lambda n I)w = X^\top y \in \mathbb{R}^p \tag{6}$$

one finds that $\widehat{w}(\lambda) = (X^\top X + \lambda n I)^{-1}X^\top y$. In order to compute $\widehat{w}(\lambda)$ there exist different methods such as :

- Method 1 : Direct inversion $O(p^3)$. In this case we have to invert a matrix of size $p \times p$.
- Method 2 : Conjugate gradient method $O(p^3)$, which is faster than method (1) in practice.

Notice (the small n large p trick): In the case where $n < p$, (6), we can look for a solution of the form $w = X^\top z$ where $z \in \mathbb{R}^n$ permits to reformulate the problem as finding : $z = (XX^\top + \lambda n I)^{-1}y$. Thus one needs to invert a matrix with a lower size $(n \times n)$.

1.2 Logistic Regression

Logistic regression is only used for classification problems. In this case, the loss function is defined as $(\forall a \in \mathbb{R}), (\forall b \in \mathbb{R}),\ l(a,b) = \log(1 + \exp(-ab))$. From first-order stationary condition

$$\nabla f(w) = 0 \quad \Longleftrightarrow \quad -\frac{1}{n}\sum_{i=1}^{n} \frac{y_i}{1+\exp(y_i x_i^\top w)} + \lambda w = 0 \tag{7}$$

it is not easy to exhibit an explicit solution for $\widehat{w}(\lambda)$. However, it is possible to find the solution by means of the following iterations $w_{t+1} = w_t - \eta_t \nabla f(w_t)$. Indeed, for this kind of problem (minimization of a strongly convex function) the gradient descent is very fast.

Theorem 1.2. *If we choose $\eta_t = \frac{1}{L+\lambda}$ where L is the Lipschitz constant of R, then*

$$f(w_t) - \min_w f(w) \leq \left(\frac{L-\lambda}{L+\lambda}\right)^t C \quad \text{, where } C \text{ is a constant.} \tag{8}$$

The second method one can use is called the Newton method.

$$f(w) = f(w_t) + \nabla f(w_t)^\top (w - w_t) + \frac{1}{2}(w-w_t)^\top \nabla^2 f(w_t)(w-w_t) + o\left(\|w-w_t\|_2^2\right) \tag{9}$$

The Newton method consist of finding a direction that minimizes the quadratic approximation, and make a step into that direction: $z_t = w_t - \eta_t \left(\nabla^2 f(w_t)\right)^{-1} \nabla f(w_t)$. The number of iterations required by the Newton method is faster than for the gradient descent method, but each iteration is more costly.

Notice (probabilistic interpretation of logistic regression): For $\mathbb{P}[y|x] = \frac{\exp(yw^\top x)}{\exp(w^\top x)+\exp(-w^\top x)}$ and assuming that (y_i, x_i) are i.i.d, then :

$$\max_w \mathbb{P}[y_1, \cdots, y_n | x_1, \cdots, x_n] \quad \Longleftrightarrow \quad \min_w - \log\left(\mathbb{P}[y_1, \cdots, y_n | x_1, \cdots, x_n]\right) \tag{10}$$

$$\max_w \Pi_{i=1}^n \mathbb{P}[y_i|x_i] \quad \Longleftrightarrow \quad \min_w \sum_{i=1}^n \log(1+\exp(-y_i x_i^\top w)) \tag{11}$$

1.3 Support Vector Machine

In this case, the loss function is the Hinge loss function defined as $(\forall a \in \mathbb{R}), (\forall b \in \mathbb{R}),\ l(a,b) = \max(0, 1-ab)$ which is convex but non smooth. The problem consists in finding the minimizer $\widehat{w}$:

$$\min_{w\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \max(0, 1-y_i x_i^\top w) + \frac{\lambda}{2}\|w\|_2^2 \tag{12}$$

By using slack variables ζ, it can be recast into :

$$\min_{w\in\mathbb{R}^p, \zeta\in\mathbb{R}^n} \frac{1}{n}\sum_{i=1}^n \zeta_i + \frac{\lambda}{2}\|w\|_2^2 \quad \text{s.t.} \quad \begin{cases} \zeta_i \geq 0 \\ \zeta_i \geq 1 - y_i x_i^\top w \end{cases} \tag{13}$$

which is called a "quadratic program" (minimizing a quadratic function under linear constraints), for which efficient solvers exists.

1.4 Kernels

$\mathcal{H}$ is an Hilbert space representing a class of functions $f : \chi \to \mathbb{R}$. In this case, for the purpose of classification, we minimize over a space of functions $\mathcal{H}$.

$$\min_{f\in\mathcal{H}} \frac{1}{n} l(y_i; f(x_i)) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2 \tag{14}$$

Example 1.3. *Linear kernel $K(x,x') = x^\top x'$, $\chi = \mathbb{R}^p$*
$\forall w \in \mathbb{R}^p$: $Kw : x \to w^\top x \in \mathcal{H}$.
and $\mathcal{H}$ is the space of linear functions.

Theorem 1.4. *Representer theorem*
Let us define the subspace $\mathcal{H}' = \{x \to \sum_{i=1}^n \alpha_i K(x_i,x), \alpha \in \mathbb{R}^n\} \subseteq \mathcal{H}$.
Then, all f solutions of (14) are subject to $f \in \mathcal{H}'$.

Proof. The proof rely on the following observation :

$$\forall f \in \mathcal{H}, \quad f = f'' + f^\perp \quad \text{where} \quad \begin{cases} f'' \in \mathcal{H}', \\ f^\perp \in \mathcal{H}'^\perp. \end{cases}$$

Then, $\|f\|^2_\mathcal{H} = \|f''\|^2_\mathcal{H} + \|f^\perp\|^2_\mathcal{H}$. $f(x_i) = \underbrace{\langle f'', Kx_i\rangle}_{=f''(x_i)} + \underbrace{\langle f^\perp, Kx_i\rangle}_{=0}$. □

This theorem is interesting because we just need to find the set of $\alpha \in \mathbb{R}^n$.

- In the case of the ridge regression :

$$\min_{f\in\mathcal{H}'} \frac{1}{n}\sum_{i=1}^n \frac{1}{2}(y_i - f(x_i)) + \frac{\lambda}{2}\|f\|^2_{\mathcal{H}'} \tag{15}$$

  Let us choose $\alpha \in \mathbb{R}^n$:

$$\begin{aligned} f(x_i) &= \langle f, Kx_i\rangle & (16)\\ &= \langle \sum_{j=1} \alpha_j Kx_j, Kx_i\rangle & (17)\\ &= \sum_{j=1}^n \alpha_j \langle Kx_j, Kx_i\rangle = [K\alpha]_i & (18)\end{aligned}$$

  In the same way, one can show that $\langle f, f\rangle = \alpha^\top K\alpha$. Then the minimization problem can be reformulate as follow :

$$\min_{\alpha\in\mathbb{R}^n} \frac{1}{2n}\|Y - K\alpha\|^2_2 + \frac{\lambda}{2}\alpha^\top K\alpha \tag{19}$$

  From the first order stationary condition, one finds that the solution of this problem is $\widehat{\alpha}(\lambda) = (K + \lambda n \mathbb{1})^{-1} Y$. Note that with a linear kernel $K = X^\top X$, we find back the "small n, large p trick"

- In the case of SVM :

$$\min_{f\in\mathcal{H}',\zeta\in\mathbb{R}^n} \frac{1}{n}\sum_{i=1}^n \zeta_i + \frac{\lambda}{2}\|f\|^2_{\mathcal{H}'} \quad \text{s.t.} \quad \begin{cases} \zeta_i \geq 0 \\ \zeta_i \geq 1 - y_i f(x_i) \end{cases} \tag{20}$$

  which can be recast into :

$$\min_{\alpha\in\mathbb{R}^n,\zeta\in\mathbb{R}^n} \frac{1}{n}\sum_{i=1}^n \zeta_i + \frac{\lambda}{2}\alpha^\top K\alpha \quad \text{s.t.} \quad \begin{cases} \zeta_i \geq 0 \\ \zeta_i \geq 1 - y_i[K\alpha]_i \end{cases} \tag{21}$$

  which can again be solved using quadratic programming.

2 Cross Validation

Given a model, how do we estimate the prediction error ? This is not something obvious. In addition, how to choose the regularization parameter λ ?

$$\widehat{w}(\lambda) = \arg\min_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} l(y_i, \underbrace{w^\top x_i}_{(f(x_i)}) + \frac{\lambda}{2} \underbrace{\|w\|_2^2}_{(\|f\|_{\mathcal{H}}^2)} \tag{22}$$

Question : $E_n(\lambda) = \mathbb{E}_{(y,x)}[l(y, \widehat{w}(\lambda)^\top x)]$?

2.1 Case with lots of data

We can cut the set of data into two parts : a training sample T and a validation sample V. [ref, p220, fig7.1]. Then we find $\widehat{w}$ by minimizing (22) over T and we choose λ which minimize the validation error.

2.2 Bias-variance decomposition

$y = f(x) + \epsilon$. Estimator $\widehat{f}$ issued from training data. Given some data x_0, and calling T the training data :

$$\begin{aligned} E_n(x_0) &= \mathbb{E}_T\left[(y_0 - \widehat{f}(x_0))^2\right] && (23)\\ &= \mathbb{E}_T\left[\left(\epsilon + f(x_0) - \widehat{f}(x_0)\right)^2\right] && (24)\\ &= \mathbb{E}_T[\epsilon^2] + \mathbb{E}_T\left[(f(x_0) - \widehat{f}(x_0))^2\right] && (25)\\ &= \sigma_\epsilon^2 + \underbrace{\mathbb{E}_T\left[(f(x_0) - \mathbb{E}_T[\widehat{f}(x_0)])^2\right]}_{\text{Biais squared}} + \underbrace{\mathbb{E}_T\left[\left(\widehat{f}(x_0) - \mathbb{E}_T\widehat{f}(x_0)\right)^2\right]}_{\text{Var}(\widehat{f}(x_0))} && (26) \end{aligned}$$

Some intuition is that a highly regularized model has low variance, but can have a large bias: for example, when λ goes to infinity, $\widehat{w}(\lambda)$ will always be close to zero. On the other hand, a low regularization can lead to low bias, but large variance. The goal of cross-validation is to find a λ which is a good trade-off.

2.3 Cross-validation

K-folds cross-validation.

1. Compute $\text{CV}(\lambda) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_{\text{Val}}} \sum_{i \in \text{Val}(k)} l(y_i, \widehat{f}^{\setminus k}(x_i))$.
2. Find $\widehat{\lambda} = \arg\min \text{CV}(\lambda)$
3. Learn $\widehat{f}_{\widehat{\lambda}}$ on the full training data
4. Test $E = \frac{1}{n_{\text{test}}} \sum_{i \in \text{Test}} L(y_i, \widehat{f}_{\widehat{\lambda}}(x_i))$

$K = 5$ or $K = 10$ are often used in practice.

3 Nearest Neighbors

M different neighbors. Training data $(x_i, y_i)_{i=1,\cdots,n}$ and $y_i \in \{1, \cdots, M\}$. Given a new test point x, how do I classify it ?

$$\widehat{y}_{\text{NN}} = \text{label}\left(\arg\min_{i=1,\cdots,n} d(x_i, x)\right) \tag{27}$$

K-Nearest neighbors [ref p466].

Theorem 3.1. *Correr and Hart (1967)*
"Asymptotically, the error rate of 1-NN is never more than twice the Bayes error rate".

We are going to give a sketch of the proof, by making simplifying assumptions.

Definition 3.2. *Bayes Estimator*

$$\widehat{y}_{Bayes}(x) = \arg\max_{y\in\{1,\cdots,M\}} \mathbb{P}[Y = y|X = x] \tag{28}$$

Note that the Bayes estimator does not exist in realistic setting since the conditional probability $\mathbb{P}[Y = y|X = x]$ is unknown. This is an "ideal" classifier, used for theoretical purposes.

Definition 3.3. *Bayes error*
Given a data point x, and a label $Y(x)$ drawn according to the conditional probability $\mathbb{P}[Y|X = x]$,

$$E_{Bayes}(x) = \mathbb{P}\left[Y \neq \widehat{y}_{Bayes}(x)\right], \tag{29}$$

Note that in this definition, only the label $Y(x)$ is a random variable, with $\mathbb{P}[Y(x) = y] = \mathbb{P}[Y = y|X = x]$. First let us show that the error rate of the nearest neighbors classifier is lower bounded by the Bayes error rate.

$$\mathbb{P}[Y(x) \neq \widehat{y}_{\text{Bayes}}(x)] = 1 - \mathbb{P}[Y(x) = \widehat{y}_{\text{Bayes}}(x)] \tag{30}$$

Assume that we live in a ideal world, where there exist some $\tilde{x}$ in the training set such that $x = \tilde{x}$, associated to a label $\tilde{Y}(\tilde{x})$, which is drawn according to $\mathbb{P}[Y|X = x]$. It is important here to notice that $Y(x)$ and $\tilde{Y}(\tilde{x})$ are two independent random variables identically distributed. Even though, they correspond to the same data point $x = \tilde{x}$, they do not necessarily have the same value! Then, we have that $\widehat{y}_{NN}(x) = \tilde{Y}(\tilde{x})$ is a random variable drawn according to $\mathbb{P}[Y|X = x]$.

$$\begin{aligned}
\mathbb{P}\Big[\widehat{y}_{NN}(x) \neq Y(x)\Big] &= \sum_{j=1}^{M} \mathbb{P}\Big[Y(x) = y_j, \widehat{y}_{NN}(x) \neq y_j x\Big] && (31)\\
&= \sum_{j=1}^{M} \mathbb{P}\big[Y(x) = y_j\big] \underbrace{\mathbb{P}\Big[\widehat{y}_{NN}(x) \neq y_j\Big]}_{1-\mathbb{P}\left[Y(x)=y_j\right]} && (32)\\
&\geq \sum_{j=1}^{M} \mathbb{P}\big[Y(x) = y_j\big]\left(1 - \mathbb{P}\big[Y(x) = \widehat{y}_{\text{Bayes}}(x)\big]\right) && (33)\\
&= 1 - \mathbb{P}\big[Y(x) = \widehat{y}_{\text{Bayes}}(x)\big] = E_{\text{Bayes}} && (34)
\end{aligned}$$

When $M = 2$, one can show that $\mathbb{P}[\widehat{y}_{NN}(x) \neq Y(x)] \leq 2E_{n,\text{Bayes}}$. Indeed,

$$\begin{aligned}
\mathbb{P}\Big[\widehat{y}_{NN}(x) \neq Y(x)\Big] &= \sum_{j=1}^{2} \mathbb{P}\big[Y(x) = y_j\big](1 - \mathbb{P}\big[Y(x) = y_j\big]) && (35)\\
&= 2\mathbb{P}\big[\widehat{y}_{\text{Bayes}}(x) = y_j\big](1 - \mathbb{P}\big[\widehat{y}_{\text{Bayes}}(x) = y_j\big]) && (36)\\
&\leq 2(1 - \mathbb{P}\big[\widehat{y}_{\text{Bayes}}(x) = y_j\big]) = 2E_{\text{Bayes}} && (37)
\end{aligned}$$

Let us now treat the case $M > 2$. To simplify the notation, we will write $P_j = \mathbb{P}\left[Y = y_j | X = x\right]$ and $P^* = \mathbb{P}\left[Y = y_{\text{Bayes}}(x) | X = x\right]$.

$$\begin{aligned}
\mathbb{P}\Big[\widehat{y}_{NN}(x) \neq Y(x)\Big] &= \sum_{j=1}^{n} P_j(1 - P_j) && (38)\\
&= P^*(1 - P^*) + \sum_{j \neq j^*} P_j(1 - P_j) && (39)\\
&= P^*(1 - P^*) + (1 - P^*) - \sum_{j \neq j^*} P_j^2 && (40)\\
&= 2(1 - P^*) - (1 - P^*)^2 - \sum_{j \neq j^*} P_j^2 && (41)
\end{aligned}$$

If one notices that

$$\left(\sum_{j \neq j^*} P_j\right)^2 \leq \sum_{j \neq j^*} P_j^2 (K - 1) \tag{42}$$

Then :

$$\sum_{j=1}^{n} P_j(1 - P_j) \leq 2(1 - P^*) - \frac{K}{K - 1}(1 - P^*)^2 \tag{43}$$

4 LASSO

$$\min_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} l(y_i, x_i^\top w) + \frac{\lambda}{2} \|w\|_2^2 \tag{44}$$

$$\min_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} l(y_i, x_i^\top w) \quad \text{s.t.} \quad \|w\|_2^2 \leq T \tag{45}$$

Assume that the "true" w is sparse, meaning that it has a lots of zeros. One way to introduce sparsity would be to minimize :

$$\min_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} l(y_i, x_i^\top w) + \frac{\lambda}{2} \Omega(w) \tag{46}$$

If we choose $\Omega(w) = \#\{w_i \neq 0\}$ then the problem is NP-hard. If we choose $\Omega(w) = \|w\|_1$ which is a convex set, the problem is easily feasible and introduce sparsity.

$$\min_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} l(y_i, x_i^\top w) + \frac{\lambda}{2} \|w\|_1 \tag{47}$$

$$\min_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} l(y_i, x_i^\top w) \quad \text{s.t.} \quad \|w\|_1 \leq T \tag{48}$$