Paris Cité **Apprentissage statistique**
2025/2026 **Stephan Clémençon - Ekhine Irurozki**

Travaux dirigés n° 1 : Classifieur de Bayes

Stephan Clémençon <stephan.clemencon@telecom-paris.fr>
Ekhine Irurozki <irurozki@telecom-paris.fr>

On se place dans le cadre de la classification binaire. Dans tout le TD, on considère un descripteur aléatoire X à valeurs dans un espace mesurable $\mathcal{X} \subset \mathbb{R}^d$ ($d \in \mathbb{N}^*$) et un label aléatoire Y valant 0 ou 1. La distribution jointe du vecteur (X, Y) est notée P, et la fonction de régression (probabilité a posteriori)

$$\eta : x \in \mathcal{X} \mapsto \mathbb{P}(Y = 1 \mid X = x) \in [0, 1].$$

Exercice 1. On considère $\mathcal{X} = [0, 1]$ et P telle que :

- la distribution conditionnelle de X sachant Y = 0 est $P_0 = \mathcal{U}([0, \theta])$ où $\theta \in]0, 1[$,
- la distribution conditionnelle de X sachant Y = 1 est $P_1 = \mathcal{U}([0, 1])$,
- $p = \mathbb{P}(Y = 1) \in]0, 1[$.

Pour $x \in \mathcal{X}$, donner $\eta(x)$ en fonction de p et θ.

Solution :
On remarque tout d'abord que les distributions P_0 et P_1 étant Uniformes, elles admettent toutes deux des densités, notées respectivement

$$f_0 : x \in \mathbb{R} \mapsto \frac{1}{\theta}\mathbb{1}_{[0,\theta]}(x) \quad \text{et} \quad f_1 : x \in \mathbb{R} \mapsto \mathbb{1}_{[0,1]}(x).$$

Par conséquent, d'après la formule des probabilités totales, la loi non conditionnelle de X admet elle aussi une densité :

$$f : x \in \mathbb{R} \mapsto f_0(x)(1-p) + f_1(x)p = \frac{1-p}{\theta}\mathbb{1}_{[0,\theta]}(x) + p\mathbb{1}_{[0,1]}(x).$$

Soit maintenant $x \in [0, 1]$. D'après la formule de Bayes, on a

$$\eta(x) = \mathbb{P}(Y = 1 \mid X = x) = \frac{f_1(x)p}{f(x)} = \frac{p\mathbb{1}_{[0,1]}(x)}{p\mathbb{1}_{[0,1]}(x) + \frac{1-p}{\theta}\mathbb{1}_{[0,\theta]}(x)} = \begin{cases} 1 & \text{si } x \in (\theta, 1], \\ \dfrac{\theta p}{1-p+\theta p} & \text{si } x \in [0, \theta], \end{cases}$$

Lorsque $\theta = \frac{1}{2}$, on obtient

$$\eta(x) = \begin{cases} 1 & \text{si } x \in \left(\frac{1}{2}, 1\right], \\ \dfrac{p}{2-p} & \text{si } x \in \left[0, \frac{1}{2}\right], \end{cases}$$

comme illustré à la Figure 1.

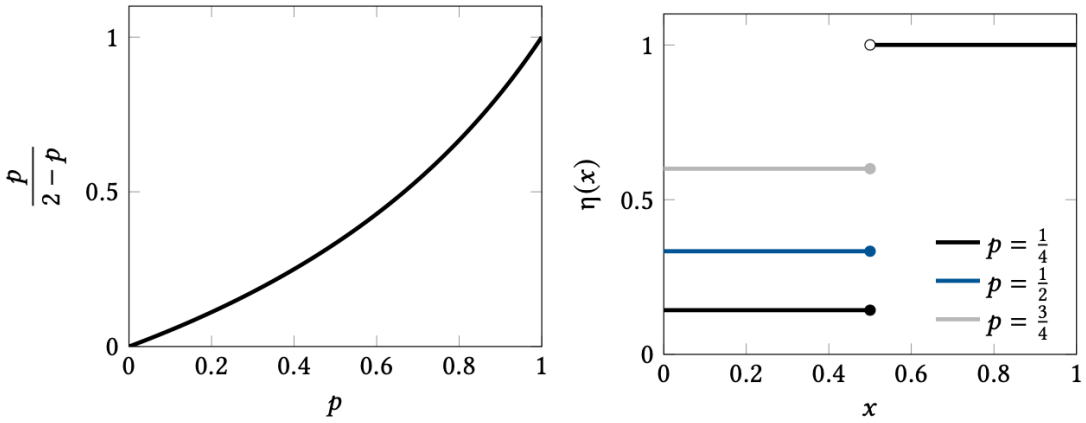

FIGURE 1 – Illustration pour $\theta = \frac{1}{2}$.

EXERCICE 2. On considère $\mathbb{P}$ telle que la distribution de X est P_X sur $\mathcal{X}$, et on note h^* le classifieur de Bayes.

1) Montrer que son risque 0-1 (sa probabilité d'erreur) vaut

$$L(h^*) = \int_{\mathcal{X}} \min\{\eta(x), 1 - \eta(x)\} \, P_X(dx).$$

2) On suppose maintenant que $X = \mathbb{R}_+$ et que, pour tout $x \in \mathbb{R}_+$, la fonction de régression vaut

$$\eta(x) = \frac{x}{x + \theta}, \quad \text{où } \theta > 0 \text{ est fixé.}$$

(i) Expliciter le classifieur de Bayes et son risque 0-1 dans ce modèle.

(ii) Calculer le risque de Bayes lorsque $P_X = \mathcal{U}([0, \alpha\theta])$ où $\alpha > 1$.

Solution :

1) Le classifieur de Bayes pour des labels valant 0 ou 1 est la fonction

$$h^* : x \in \mathcal{X} \mapsto \mathbb{1}_{\{\eta(x) > \frac{1}{2}\}}.$$

On peut remarquer que la fonction η étant à valeurs dans $[0, 1]$, h^* peut être réécrit comme

$$h^* : x \in \mathcal{X} \mapsto \mathbb{1}_{\{\eta(x) > 1 - \eta(x)\}}.$$

Son risque 0-1 vaut donc :

$$\begin{aligned}
L(h^*) &= \mathbb{E}[\mathbb{1}_{\{Y \neq h^*(X)\}}] \\
&= \mathbb{E}[\mathbb{1}_{\{Y=0\}} \mathbb{1}_{\{h^*(X)=1\}} + \mathbb{1}_{\{Y=1\}} \mathbb{1}_{\{h^*(X)=0\}}] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\{Y=0\}} \mathbb{1}_{\{h^*(X)=1\}} + \mathbb{1}_{\{Y=1\}} \mathbb{1}_{\{h^*(X)=0\}} \mid X\right]\right] \quad \text{(espérance totale)} \\
&= \mathbb{E}\left[(1 - \eta(X)) \mathbb{1}_{\{h^*(X)=1\}} + \eta(X) \mathbb{1}_{\{h^*(X)=0\}}\right] \\
&= \mathbb{E}\left[(1 - \eta(X)) \mathbb{1}_{\{\eta(X) > 1 - \eta(X)\}} + \eta(X) \mathbb{1}_{\{\eta(X) \leq 1 - \eta(X)\}}\right] \\
&= \mathbb{E}[\min(\eta(X), 1 - \eta(X))] \\
&= \int_{\mathcal{X}} \min(\eta(x), 1 - \eta(x)) \, dP_X(x).
\end{aligned}$$

2) (a) Soit $x \in \mathbb{R}^+$, alors (cf. Figure 2)

$$\frac{1}{\eta(x)} > 2 \quad \Leftrightarrow \quad \frac{x}{x+\theta} > \frac{1}{2} \quad \Leftrightarrow \quad x > \frac{x+\theta}{2} \quad \Leftrightarrow \quad x > \theta.$$

Ainsi, $h^*(x) = \mathbb{1}_{\{x>\theta\}}$ et d'après la question précédente

$$\begin{aligned}
\mathrm{L}(h^*) &= \int_0^{+\infty} \min(\eta(x), 1-\eta(x))\, \mathrm{P_X}(\mathrm{d}x) \\
&= \int_0^{+\infty} \min\left(\frac{x}{x+\theta}, \frac{\theta}{x+\theta}\right) \mathrm{P_X}(\mathrm{d}x) \\
&= \int_0^{\theta} \frac{x}{x+\theta} \mathrm{P_X}(\mathrm{d}x) + \int_\theta^{+\infty} \frac{\theta}{x+\theta} \mathrm{P_X}(\mathrm{d}x) \\
&= \int_0^{\theta} \left(1 - \frac{\theta}{x+\theta}\right) \mathrm{P_X}(\mathrm{d}x) + \int_\theta^{+\infty} \frac{\theta}{x+\theta} \mathrm{P_X}(\mathrm{d}x) \\
&= \mathbb{P}(\mathrm{X} \le \theta) + \int_\theta^{+\infty} \frac{\theta}{x+\theta} \mathrm{P_X}(\mathrm{d}x) - \int_0^{\theta} \frac{\theta}{x+\theta} \mathrm{P_X}(\mathrm{d}x).
\end{aligned}$$

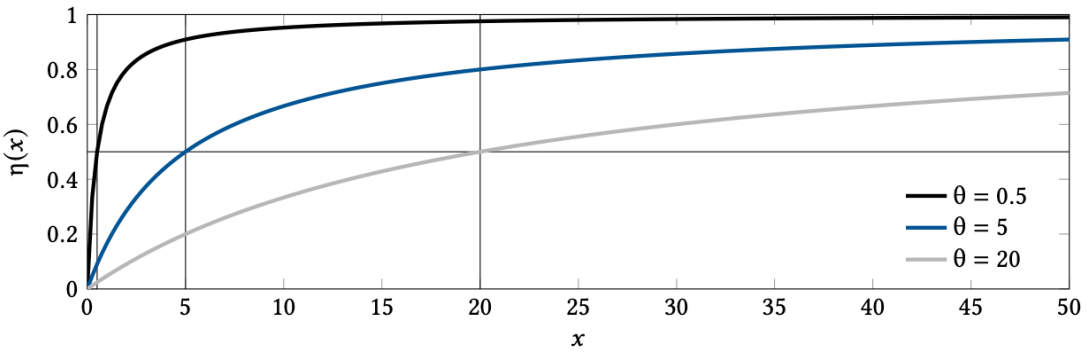

FIGURE 2 – Tracé de la fonction de régression pour différentes valeurs de θ.

(b) Lorsque $\mathrm{P_X}$ est une loi uniforme sur $[0, \alpha\theta]$ avec $\alpha > 1$, on obtient

$$\begin{aligned}
\mathrm{L}(h^*) &= \frac{\theta}{\alpha\theta} + \int_\theta^{+\infty} \frac{\theta}{x+\theta} \frac{1}{\alpha\theta} \mathbb{1}_{[0,\alpha\theta]}(x)\, \mathrm{d}x - \int_0^{\theta} \frac{\theta}{x+\theta} \frac{1}{\alpha\theta} \mathbb{1}_{[0,\alpha\theta]}(x)\, \mathrm{d}x \\
&= \frac{1}{\alpha}\left(1 + \int_\theta^{\alpha\theta} \frac{1}{x+\theta}\, \mathrm{d}x - \int_0^{\theta} \frac{1}{x+\theta}\, \mathrm{d}x\right) \\
&= \frac{1}{\alpha}\left(1 + \int_{2\theta}^{(1+\alpha)\theta} \frac{1}{x}\, \mathrm{d}x - \int_\theta^{2\theta} \frac{1}{x}\, \mathrm{d}x\right) \\
&= \frac{1}{\alpha}(1 + \ln((\alpha+1)\theta) - \ln(2\theta) - \ln(2\theta) + \ln(\theta)) \\
&= \frac{1}{\alpha} \ln\left(e\frac{\alpha+1}{4}\right).
\end{aligned}$$

Cette fonction de α est illustrée à la Figure 3.

Interprétation Posons $p := \mathbb{P}(\mathrm{Y} = 1)$. On remarque que, sous les hypothèses de la question 2.b, on peut réécrire le risque de Bayes comme suit :

$$\mathrm{L}(h^*) = \mathbb{P}(\mathrm{X} \le \theta | \mathrm{Y} = 1)p + \mathbb{P}(\mathrm{X} > \theta | \mathrm{Y} = 0)(1-p).$$

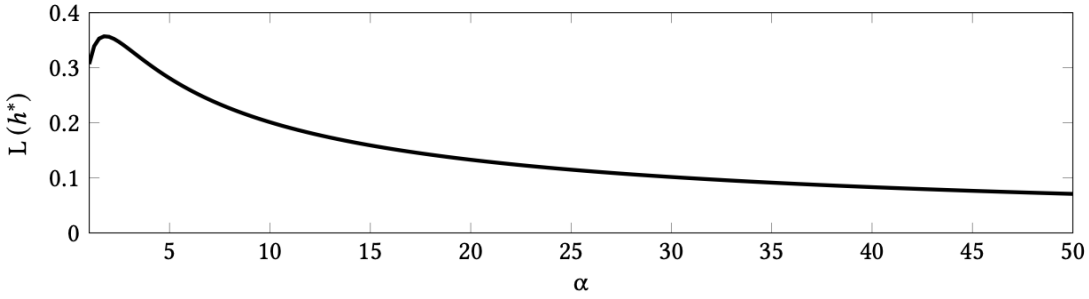

FIGURE 3 – Tracé de la fonction de régression pour différentes valeurs de θ.

On peut expliciter p :

$$p := \mathbb{P}(Y = 1) = \int_X \eta(x) P_X(dx) = \frac{1}{\alpha\theta} \int_0^{\alpha\theta} \frac{x}{x+\theta} dx = 1 - \frac{1}{\alpha} \ln(\alpha + 1).$$

Par ailleurs, en notant f la densité de X, d'après la formule de Bayes, X sachant Y = 1 possède une densité f_1 valant pour tout $x \in \mathbb{R}$:

$$f_1(x) = \frac{\eta(x) f(x)}{p} = \frac{x \mathbb{1}_{[0,\alpha\theta]}(x)}{(x+\theta)\theta(\alpha - \ln(\alpha+1))}.$$

et X sachant Y = 0 possède une densité f_0 valant pour tout $x \in \mathbb{R}$.

$$f_0(x) = \frac{(1-\eta(x)) f(x)}{1-p} = \frac{x \mathbb{1}_{[0,\alpha\theta]}(x)}{(x+\theta)(\ln(\alpha+1))}.$$

Cela nous permet de représenter graphiquement le risque de Bayes en fonction des lois du descripteur X dans chacune des classes déterminées par le label Y, comme à la Figure 4.

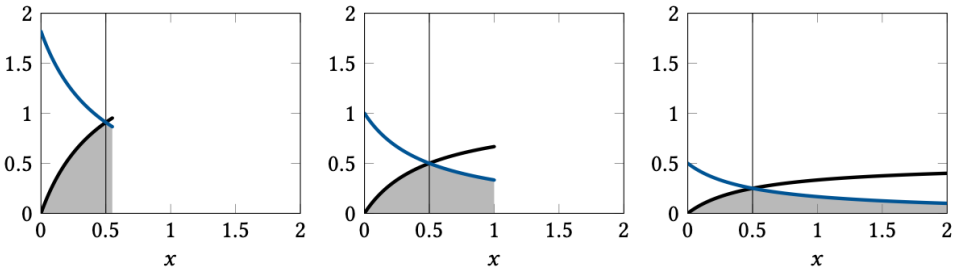

FIGURE 4 – Représentation graphique du risque de Bayes (aire grisée) selon pf_1 (courbe en noir) et $(1-p)f_0$ (courbe en bleu) pour θ = 0.5 (trait vertical) et α = 1.1 (gauche), α = 2 (milieu), α = 4 (droite), quand $P_X = \mathcal{U}([0, \alpha\theta])$.

On retrouve bien dans l'aire grisée de la Figure 4 l'expression intégrale de la question 1 :

$$L(h^*) = \int_0^{+\infty} \min(\eta(x), 1-\eta(x)) f(x)\, dx = \int_0^{+\infty} \min(pf_1(x), (1-p)f_0(x))\, dx.$$

Remarque. Pour se convaincre que le classifieur de Bayes pour des labels valant 0 ou 1 est bien celui donné, on peut procéder comme suit. Dans le cours, il a été vu que pour des labels valant -1 ou 1, le classifieur de Bayes n'était autre que la fonction

$$g : x \in \mathcal{X} \mapsto 2\mathbb{1}_{\{\eta(x) > \frac{1}{2}\}} - 1,$$

qui minimise le risque 0-1 parmi tous les classifieurs $g : X \to \{-1, 1\}$ possibles. On remarque que la variable aléatoire $Z := 2Y - 1$ prend ses valeurs dans {-1, 1}, et $\mathbb{P}(Z = 1) = \mathbb{P}(Y = 1)$. Alors

$$\mathbb{E}\{\mathbb{1}_{\{Y \neq h(X)\}}\} = \mathbb{E}\{\mathbb{1}_{\{Z \neq 2h(X)-1\}}\}$$

est bien minimal pour h^* tel que $2h^* - 1 = g^*$.

Exercice 3. Soient des poids $\omega(0), \omega(1) \geq 0$ tels que $\omega(0) + \omega(1) = 1$. On considère le risque de classification pondéré :

$$L_\omega(g) = \mathbb{E}\left(2\omega(Y) \cdot \mathbb{1}_{\{Y \neq g(X)\}}\right), \qquad g : \mathcal{X} \to \{0, 1\}.$$

Donner le classifieur de Bayes et le risque de Bayes pour ce critère. Quel est l'intérêt de considérer un tel critère ?

Solution :
Le classifieur de Bayes pour le risque de classification pondéré est par définition l'application $g^* : \mathcal{X} \to \{0, 1\}$ telle que pour tout classifieur $g : \mathcal{X} \to \{0, 1\}$,

$$R_\omega(g) := L_\omega(g) - L_\omega(g^*) \geq 0.$$

Soit $g : X \to \{0, 1\}$ un classifieur quelconque. Alors

$$\begin{aligned}
L_\omega(g) &= \mathbb{E}\left[2\omega_Y \mathbb{1}_{\{Y \neq g(X)\}}\right] \\
&= 2\mathbb{E}\left[\omega_1 \mathbb{1}_{\{Y=1\}} \mathbb{1}_{\{g(X)=0\}}\right] + \omega_0 \mathbb{E}\left[\mathbb{1}_{\{Y=0\}} \mathbb{1}_{\{g(X)=1\}}\right] \\
&= 2\mathbb{E}\left[\omega_1 \eta(X) \mathbb{1}_{\{g(X)=0\}}\right] + \omega_0 \mathbb{E}\left[(1 - \eta(X)) \mathbb{1}_{\{g(X)=1\}}\right] \quad \text{(espérance totale)} \\
&= 2\mathbb{E}\left[\omega_1 \eta(X)\right] + \left(\omega_0(1 - \eta(X)) - \omega_1 \eta(X)\right) \mathbb{1}_{\{g(X)=1\}} \\
&\quad \left(since \mathbb{1}_{\{g(X)=0\}} = 1 - \mathbb{1}_{\{g(X)=1\}} \text{ p.s.}\right) \\
&= 2\mathbb{E}\left[\omega_1 \eta(X)\right] + (\omega_0 - \eta(X)) \mathbb{1}_{\{g(X)=1\}}, \quad (since \omega_1 + \omega_0 = 1)
\end{aligned}$$

d'où

$$R_\omega(g) = 2\mathbb{E}\left[\omega_0 - \eta(X)\right]\left(\mathbb{1}_{\{g(X)=1\}} - \mathbb{1}_{\{g^*(X)=1\}}\right).$$

Or pour tout $x \in \mathcal{X}$ on a $(\omega_0 - \eta(x))(\mathbb{1}_{\{g(x)=1\}} - \mathbb{1}_{\{g^*(x)=1\}}) \geq 0$ ssi

$$\begin{aligned}
\omega_0 - \eta(x) &\geq 0, \\
\mathbb{1}_{\{g(x)=1\}} - \mathbb{1}_{\{g^*(x)=1\}} &\geq 0,
\end{aligned}$$

ou

$$\begin{aligned}
\omega_0 - \eta(x) &\leq 0, \\
\mathbb{1}_{\{g(x)=1\}} - \mathbb{1}_{\{g^*(x)=1\}} &\leq 0.
\end{aligned}$$

La fonction $g^* : x \in \mathcal{X} \mapsto \mathbb{1}_{\{\eta(x) > \omega_0\}}$ satisfait cette condition sur tout $\mathcal{X}$, et en intégrant elle permet d'obtenir $R_\omega(g) \geq 0$ quel que soit g.

Son risque pondéré vaut alors

$$\begin{aligned}
L_\omega(g^*) &= 2\mathbb{E}\left[\omega_1 \eta(X) \mathbb{1}_{\{\eta(X) \leq \omega_0\}} + \omega_0(1 - \eta(X)) \mathbb{1}_{\{\eta(X) > \omega_0\}}\right] \\
&= 2\mathbb{E}\left[\omega_1(\eta(X) + \omega_0 - \omega_0) \mathbb{1}_{\{\eta(X) \leq \omega_0\}} + \omega_0(1 - \omega_0 - \eta(X) + \omega_0) \mathbb{1}_{\{\eta(X) > \omega_0\}}\right] \\
&= 2\omega_0\omega_1 - 2\mathbb{E}\left[\left|\eta(X) - \omega_0\right| \left(\omega_1 \mathbb{1}_{\{\eta(X) \leq \omega_0\}} + \omega_0 \mathbb{1}_{\{\eta(X) > \omega_0\}}\right)\right].
\end{aligned}$$

On retrouve bien le risque de Bayes classique $\frac{1}{2} - \mathbb{E}\left[\left|\eta(X) - \frac{1}{2}\right|\right]$ lorsque $\omega_0 = \omega_1 = \frac{1}{2}$.

En remarquant que pour tout $x \in \mathcal{X}$ on a $\eta(x) \leq \omega_0$ ssi $\omega_1\eta(x) \leq \omega_0(1-\eta(x))$, on peut aussi écrire le risque pondéré de g^* comme

$$L_\omega(g^*) = \mathbb{E}\left[\min\left(\omega_1\eta(X), \omega_0(1-\eta(X))\right)\right].$$

L'intérêt de ce critère est de se prémunir plus fortement contre un type d'erreur en particulier (faux positifs ou faux négatifs). Par exemple, pour un test de grossesse, il est plus important de garantir un faible taux de faux négatifs que de faux positifs, puisqu'en cas de test négatif, une femme n'ira pas consulter de médecin.

Exercice 4. On considère $X = (T, U, V)$ où T, U, V sont des variables aléatoires réelles i.i.d. de loi exponentielle standard. On pose $Y = \mathbb{1}_{\{T+U+V<\theta\}}$ où $\theta \in \mathbb{R}_+^*$ est fixé.

1) (i) Rappeler la densité f_1 et la fonction de répartition F_1 d'une loi exponentielle standard.
   (ii) Calculer la densité f_2 et la fonction de répartition F_2 de la variable aléatoire $T + U$.
   (iii) Calculer la densité f_3 et la fonction de répartition F_3 de la variable aléatoire $T + U + V$.
2) Calculer le classifieur de Bayes $(t, u) \in \mathbb{R}_+^2 \mapsto g_1^*(t, u) \in \{0, 1\}$ lorsque V n'est pas observée. Calculer le risque 0-1 associé à ce classifieur. En donner une approximation numérique lorsque $\theta = 9$.
3) On suppose à présent que seule T est observée. Reprendre les calculs précédents puis comparer les risques 0-1 obtenus lorsque $\theta = 9$.
4) Proposer un classifieur lorsque X n'a aucune composante observée. Calculer son risque 0-1 et en donner une approximation numérique lorsque $\theta = 9$. Qu'en concluez-vous ?

Conseils bibliographiques

Vous trouverez ci-dessous quelques points d'entrée utiles pour l'apprentissage automatique :

- Théorique et porté sur les aspects probabilistes : [DGL97]
- Utilitaire et porté sur les aspects pratiques : [HTF13]
- Livre récent porté essentiellement sur l'aspect optimisation : [SSBD14] (et du même auteur sur l'apprentissage en ligne [SS12])
- Méthodes Bayésiennes et modèles graphiques : [MB12]

Références

[DGL97] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Stochastic Modelling and Applied Probability. Springer New York, 1997. 6

[HTF13] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning : Data Mining, Inference, and Prediction.* Springer Series in Statistics. Springer New York, 2013. 6

[MB12] K.P. Murphy and F. Bach. *Machine Learning : A Probabilistic Perspective.* Adaptive Computation and Machine Learning Series. MIT Press, 2012. 6

[SS12] S. Shalev-Shwartz. *Online Learning and Online Convex Optimization.* Foundations and Trends in Machine Learning Series. Now Publishers, 2012. 6

[SSBD14] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning : From Theory to Algorithms.* Cambridge University Press, 2014. 6