Introduction to Machine Learning
Answers to Exercise 2 - Classification & Overfitting

Jingtao Min

March 31, 2022

1 True-False Classification with Asymmetric Losses

The numbers of true/false negatives/positives are as follows:

Table 1: Classification result distribution for classifier A, B

	$y=-1$	$y=+1$
$\hat{y}=-1$	TN = 4	FN = 2
$\hat{y}=+1$	FP = 2	TP = 3

(a) False Positive Rate (FPR) of classifier A.

(b) False Discovery Rate (FDR) of classifier A.

(c) Precision of classifier A.

(d) Recall of classifier A.

$$\begin{aligned} \text{FPR} &\sim P(\hat{y}=+1|y=-1) = \frac{\text{FP}}{\text{FP}+\text{TN}} = 1/3 \\ \text{FDR} &\sim P(y=-1|\hat{y}=+1) = \frac{\text{FP}}{\text{FP}+\text{TP}} = 2/5 \\ \text{Precision} &\sim P(y=+1|\hat{y}=+1) = \frac{\text{TP}}{\text{TP}+\text{FP}} = 3/5 \\ \text{Recall} &\sim P(\hat{y}=+1|y=+1) = \frac{\text{TP}}{\text{TP}+\text{FN}} = 3/5 \end{aligned} \tag{1}$$

(e) F1-score is the harmonic average of precision and mean:

$$\text{F1} = \frac{2}{5/3+5/3} = \frac{3}{5} \tag{2}$$

(f)
- F1-score basically yield equal weight to recall and precision, and is not the silver-bullet for evaluating classifiers. For instance, when false positive is much more serious than false negative, one may want to evaluate these two errors asymmetrically. In these cases precision is far more important than recall.
- On the contrary. Since the dataset is not linearly separable (and A does not linearly separate the positive and negative results), it can only be achieved by using soft-margin SVM.
- Seems true, if we assume 'adversarial perturbations' refer to small perturbations in the feature space. A is more robust because the correctly classified points are all distant from the boundary.

(g) $\hat{f}$ for E may be independent of y, because it may be a random classifier.

(h) $\{C, D\}$ may contain the optimal classifier.

2 Ridge Regression

The loss function for Ridge regression:

$$\arg\min_{\mathbf{w}} L_{\text{Ridge}}(\mathbf{w}) = \arg\min_{\mathbf{w}} \left[\sum_{i=1}^{n} \left(y_i - \mathbf{w}^T \mathbf{x}_i \right)^2 + \lambda \left\| \mathbf{w} \right\|^2 \right] = \arg\min_{\mathbf{w}} \left(\left\| \mathbf{y} - \mathbf{X}\mathbf{w} \right\|^2 + \lambda \left\| \mathbf{w} \right\|^2 \right) \tag{3}$$

(a) The Hessian of Ridge objective function is given by:

$$\mathbf{H}_{\text{Ridge}} = \mathbf{X}^T \mathbf{X} + \lambda \mathbf{I} \qquad (\lambda > 0) \tag{4}$$

it is evidently positive definite, let alone positive semi-definite.

(b) For a *strongly-convex* function f defined on $\mathbb{R}^d$, it must have minimizer. This can be proven as follows. Let x be any point in the domain, consider the convexity defined at $\mathbf{0}$ (any arbitrary reference point works just as fine):

$$\begin{aligned} &f(\mathbf{x}) \geq f(\mathbf{0}) + \nabla f(\mathbf{0})^T \mathbf{x} + \frac{\alpha}{2} \left\| \mathbf{x} \right\|^2 \geq f(\mathbf{0}) - \left\| \nabla f(\mathbf{0}) \right\| \left\| \mathbf{x} \right\| + \frac{\alpha}{2} \left\| \mathbf{x} \right\|^2 \\ &f(\mathbf{x}) \to +\infty \left(\left\| \mathbf{x} \right\| \to +\infty \right) \quad \Longleftrightarrow \quad f(\mathbf{x}) \to +\infty \left(\mathbf{x} \to \infty \right) \end{aligned} \tag{5}$$

Therefore function f has lower bound on $\mathbb{R}^d$, and hence has minimizers. Let $\mathbf{x}_0^*$ be one minimizer of the function, and due to its being continuously differentiable, it must satisfy first-order optimality condition:

$$\nabla f(\mathbf{x}_0^*) = \mathbf{0} \tag{6}$$

Combining this condition with the strong convexity, one can establish a lower bound for any point in the domain:

$$\begin{aligned} &f(\mathbf{x}) \geq f(\mathbf{x}_0^*) + \nabla f(\mathbf{x}_0^*)^T \left(\mathbf{x} - \mathbf{x}_0^* \right) + \frac{\alpha}{2} \left\| \mathbf{x} - \mathbf{x}_0^* \right\|^2 \\ &f(\mathbf{x}) \geq f(\mathbf{x}_0^*) + 0 + \frac{\alpha}{2} \left\| \mathbf{x} - \mathbf{x}_0^* \right\|^2 = f(\mathbf{x}_0^*) + \frac{\alpha}{2} \left\| \mathbf{x} - \mathbf{x}_0^* \right\|^2 > f(\mathbf{x}_0^*) \quad (\forall \mathbf{x} \neq \mathbf{x}_0^*) \end{aligned} \tag{7}$$

This indicates that if a point $\mathbf{x}_0^*$ is a minimizer, it will also be the unique global minimizer in $\mathbb{R}^d$.

(c) The loss function for Ridge regression is fully quadratic. It can be rewritten in the following form:

$$L_{\text{Ridge}} = \left(\mathbf{y} - \mathbf{X}\mathbf{w} \right)^T \left(\mathbf{y} - \mathbf{X}\mathbf{w} \right) + \lambda \mathbf{w}^T \mathbf{w} = \mathbf{y}^T \mathbf{y} - 2\mathbf{y}^T \mathbf{X}\mathbf{w} + \mathbf{w}^T \left(\mathbf{X}^T \mathbf{X} + \lambda I \right) \mathbf{w} \tag{8}$$

Once can quickly conclude that since $\mathbf{H}_{\text{Rodge}} \succeq \lambda \mathbf{I}$, $\lambda > 0$. the loss function is *strongly convex*. This implies that a unique minimizer, and also a global minimizer $\mathbf{w}_{\text{Ridge}}^*$ exists $\forall \mathbf{X} \in \mathbb{R}^{n \times d}$.

(d) As its name suggests, the term $\lambda \left\| \mathbf{w} \right\|^2$ 'regularizes' the topography of the original loss function. With increasing regularization strength, this term conditions the topography of the original loss function to be more like a well-conditioned, isotropic quadratic function, at the cost of losing information of $\mathbf{X}$ and $\mathbf{y}$.

When $\lambda \to +\infty$, $\mathbf{w}_{\text{Ridge}}^* \to 0$. This is the case when the topography is completely dominated by $\mathbf{I}$, and the minimum is achieved at the origin. When $\lambda \to 0$, $\mathbf{w}$ approaches the minimizer of the original loss function, but suffers from ill-conditioning of the original problem (e.g. non-uniqueness, etc.).

3 Subgradients and Lasso

Linear regression with Lasso regularization takes the form:

$$\arg\min_{\mathbf{w}\in\mathbb{R}^d}\left\{\frac{1}{2}\left\|\mathbf{y}-\mathbf{X}\mathbf{w}\right\|_2^2+\lambda\left\|\mathbf{w}\right\|_1\right\} \tag{9}$$

To facilitate analyzing the minimizer, we invoke the concept of subgradient. A subgradient of a convex function $f:\mathbb{R}^d\mapsto\mathbb{R}$ at the point $\mathbf{x}$ is a vector $\mathbf{p}$ such that:

$$f(\mathbf{z})\geq f(\mathbf{x})+\langle\mathbf{p},\mathbf{z}-\mathbf{x}\rangle\qquad\forall\mathbf{z}\in\mathbb{R}^d \tag{10}$$

The set of all subgradients at $\mathbf{x}$ is denoted by $\partial f(\mathbf{x})$. For differentiable f, $\partial f(\mathbf{x})=\{\nabla f(\mathbf{x})\}$.

(a) For $f(x)=|x|$, the subgradients at point $x=0$ is given by:

$$\nabla f(0)=\{p:|p|\leq 1\}=[-1,1] \tag{11}$$

(b) For the function $f(x)=\left\|\mathbf{x}\right\|_1=\sum_{i=1}^d x_i$, its subgradients at $\mathbf{x}=\mathbf{0}$ is given by:

$$\nabla f(\mathbf{0})=\left\{\mathbf{p}:\mathbf{p}\in\mathbb{R}^d,\quad |p_i|\leq 1\,(\forall i=1,\cdots d)\right\}=\{\mathbf{p}:\left\|\mathbf{p}\right\|_\infty\leq 1\} \tag{12}$$

The conclusion follows naturally from the fact that:

$$f(\mathbf{x})-f(\mathbf{0})=\sum_{i=1}^d|x_i|,\quad\langle\mathbf{p},\mathbf{x}\rangle=\sum_{i=1}^d p_ix_i \tag{13}$$

We know that when $|p_i|\leq 1$ for all p_i, every term in the summation satisfies the subgradient criterion; and when some $|p_k|>1$, a trial vector $\hat{\mathbf{x}}$ with $x_k=1$ and $x_i=0$ $(\forall i\neq k)$ violates the subgradient criterion.

(c) We first prove the following proposition. Let $g(\mathbf{x})$ and $h(\mathbf{x})$ be two convex functions, with subgradients ∂g and ∂h, respectively. $f=g+h$ is naturally a convex function, with subgradients ∂f. Firstly, $\forall\mathbf{p}\in\partial g(\mathbf{x})$ and $\forall\mathbf{q}\in\partial h(\mathbf{x})$, we have:

$$f(\mathbf{z})=g(\mathbf{z})+h(\mathbf{z})\geq g(\mathbf{x})+\langle\mathbf{p},\mathbf{z}-\mathbf{x}\rangle+h(\mathbf{x})+\langle\mathbf{q},\mathbf{z}-\mathbf{x}\rangle=f(\mathbf{x})+\langle\mathbf{p}+\mathbf{q},\mathbf{z}-\mathbf{x}\rangle \tag{14}$$

holds true for all $\mathbf{z}\in\mathbb{R}^d$. Therefore, $\mathbf{p}+\mathbf{q}\in\partial f(\mathbf{x})$. Its inverse proposition, i.e. if $\mathbf{s}\in\partial f$, it can be decomposed into $\mathbf{p}+\mathbf{q}$, where $\mathbf{p}\in\partial g(\mathbf{x})$ and $\mathbf{q}\in\partial h(\mathbf{x})$. I assume it is true here, but there is still some subtlety in the proof. The proof should be evident once one of g and h is differentiable at $\mathbf{x}$, which is indeed the case here.

It follows from the proposition that the subgradient of Lasso problem is:

$$\partial L_{\text{Lasso}}(\mathbf{w})=\left\{-\mathbf{X}^T(\mathbf{y}-\mathbf{X}\mathbf{w})+\lambda\mathbf{p}:\mathbf{p}\in\partial\left\|\mathbf{w}\right\|_1\right\} \tag{15}$$

Let $\mathbf{w}^*$ be a minimzer of the problem, it is sufficient and necessary that $\mathbf{0}$ is included in $\partial L_{\text{Lasso}}(\mathbf{w}^*)$. Therefore,

$$\exists\mathbf{p}\in\partial\left\|\mathbf{w}^*\right\|_1\quad\text{s.t.}\quad\mathbf{X}^T(\mathbf{y}-\mathbf{X}\mathbf{w}^*)=\lambda\mathbf{p} \tag{16}$$

(d) For 1D case, the equation is given by:

$$\sum_{i=1}^n x_i\left(y_i-x_iw\right)=\lambda p,\qquad w=\frac{\sum x_iy_i-\lambda p}{\sum x_i^2} \tag{17}$$

Note p is dependent on $\mathbf{w}^*$. When $\mathbf{w}>0$, $p=1$; when $\mathbf{w}<0$, $p=-1$. When $\mathbf{w}=0$, $p\in[-1,1]$. Summarizing different situations, one can come up with the following solution of the minimizer of the 1D Lasso problem:

$$\mathbf{w}^*=\begin{cases}\left(1-\dfrac{\lambda}{|\sum x_iy_i|}\right)\dfrac{\sum_{i=1}^n x_iy_i}{\sum_{i=1}^n x_i^2}, & \lambda<\left|\sum_{i=1}^n x_iy_i\right|\\ 0, & \lambda\geq\left|\sum_{i=1}^n x_iy_i\right|\end{cases} \tag{18}$$

4 Model selection and regularization

4.1 Validation Sets

(a) False. There seems no reason why a smaller validation set can yield a better proxy of true generalization error.

(b) False. Although it might be dependent on the specific dataset and parameterization, there is no apparent reason why a smaller training set is more likely to yield a model with lower generalization error.

(c) False. It is indeed quite probable that the model performs better on training set than on validation set, but it may not always be the case.

4.2 Cross-validation

(a) False. The aim of Leave-One-Out Cross Validation (LOOCV) is to evaluate the performance of a machine learning algorithm, in particular the robustness of certain hyperparameters. It is in general not used to estimate generalization error of a prediction model (esp. the model generated at last will be trained on a different dataset than those evaluated within LOOCV).

(b) False. Same reason above, but more apparent, since training error is not even relevant in LOOCV.

(c) True. For a deterministic learning algorithm (i.e. no randomness involved), LOOCV always yields the same result, as the training-testing split is exhaustive and therefore exactly the same at each run.

4.3 Akaike Information Criterion

The Akaike Information Criteria $\text{AIC} = 2k - 2\ln\hat{L}$, where $\hat{L}$ is the likelihood of f on D_{train}, k is a measure of the number of parameters used to fit f.

(a) False. Seems irrelevant, although models with low generalization error probably (and ideally) have low AIC scores.

(b) False. $2k$ term in AIC can be viewed as a proxy of model complexity, thus penalizing overly-complex models for data fitting.

4.4 Regularization

(a) Right. Regularization term in Ridge regression has concentric hyperspheres as its contours.

(b) Left. Regularization term in Lasso regression has 'hyper-diamonds' as its contours.

5 Support Vector Machine (SVM)

(a) For this simple dataset it is apparent that the boundary should be somewhere between B and C. The hyperplane (=straight line in this context) that maximizes the margin is the perpendicular bisector of the segment BC. Therefore its equation:

$$\begin{aligned}(x_{\mathrm{B}}-x_{\mathrm{C}})\left(x-\frac{x_{\mathrm{B}}+x_{\mathrm{C}}}{2}\right)+(y_{\mathrm{B}}-y_{\mathrm{C}})\left(y-\frac{y_{\mathrm{B}}+y_{\mathrm{C}}}{2}\right)&=0\\ (x_{\mathrm{B}}-x_{\mathrm{C}})x+(y_{\mathrm{B}}-y_{\mathrm{C}})y-\frac{1}{2}\left(x_{\mathrm{B}}^2+y_{\mathrm{B}}^2-x_{\mathrm{C}}^2-y_{\mathrm{C}}^2\right)&=0\\ -x+y-\frac{1}{2}(4+6.25-9-2.25)=-x+y+\frac{1}{2}&=0\end{aligned} \tag{19}$$

(b) Now the margin is really just given by support vectors B and C. Therefore,

$$\text{margin}=\frac{-x_{\mathrm{B}}+y_{\mathrm{B}}+1/2}{\sqrt{(-1)^2+1^2}}=\frac{-x_{\mathrm{C}}+y_{\mathrm{C}}+1/2}{\sqrt{(-1)^2+1^2}}=\frac{1}{\sqrt{2}}\approx 0.707 \tag{20}$$

(c) Graphical illustration.

(d) The objective of maximum margin classifier:

$$\begin{aligned}\hat{\mathbf{w}}&=\arg\max_{\mathbf{w}}\left\{\frac{\min_i\{y_i\langle\mathbf{w},\mathbf{x}_i\rangle\}}{\|\mathbf{w}\|}\right\}=\arg\max_{\mathbf{w}}\left\{\frac{\text{margin}(\mathbf{w})}{\|\mathbf{w}\|}\right\}\\ &=\arg\max_{\mathbf{w}}\left\{\left(\frac{\|\mathbf{w}\|}{\text{margin}(\mathbf{w})}\right)^{-1}\right\}=\arg\min_{\mathbf{w}}\left\{\frac{\|\mathbf{w}\|}{\text{margin}(\mathbf{w})}\right\}\end{aligned} \tag{21}$$

Note that this formulation only constrains the direction of $\mathbf{w}$. Any scaler factor multiplied with $\mathbf{w}$ does not change the minimizer. Therefore we can constrain the modulus of $\mathbf{w}$ so that $\text{margin}(\mathbf{w})\equiv 1$, or equivalently $\mathbf{w}:=\mathbf{w}/\text{margin}(\mathbf{w})$. Under this constraint:

$$\begin{aligned}\hat{\mathbf{w}}&=\arg\min_{\mathbf{w}}\{\|\mathbf{w}\|\} && \text{s.t.}\quad \text{margin}(\mathbf{w})=1\\ &=\arg\min_{\mathbf{w}}\left\{\tfrac{1}{2}\|\mathbf{w}\|^2\right\} && \text{s.t.}\quad y_i\langle\mathbf{x}_i,\mathbf{w}\rangle\geq 1\quad\forall i=1\cdots N\end{aligned} \tag{22}$$

In this context, we'll naturally have normalized margin

$$\widehat{\text{margin}}(\hat{\mathbf{w}})=\left|\left\langle\mathbf{x}_{\text{sv}},\frac{\hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|}\right\rangle\right|=\frac{\text{margin}(\hat{\mathbf{w}})}{\|\hat{\mathbf{w}}\|}=\frac{1}{\|\hat{\mathbf{w}}\|} \tag{23}$$