Module 7: Multiclass Linear Prediction, Generalization, and Distribution Shift

Machine Learning Course

Contents

1 Introduction

Multiclass supervised learning demands both accurate prediction and formal guarantees that the accuracy will persist on future data. This module ties *three* linear models (multiclass Perceptron, soft-margin SVM, multinomial logistic regression) to the statistical-learning framework, shows how *margin size* controls sample complexity, and discusses what happens when the train/test distributions drift apart.

2 Mathematical Framework

Let the instance space be $\mathcal{X} \subseteq \mathbb{R}^d$, label space $\mathcal{Y} = \{1, \ldots, k\}$, and P the unknown joint distribution on $\mathcal{X} \times \mathcal{Y}$. Given a training sample $S = \{(x^{(i)}, y^{(i)})\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P$, the goal is to select a hypothesis $h : \mathcal{X} \to \mathcal{Y}$ minimising the *true error*

$$\mathrm{TrueErr}(h) = \Pr_{(x,y)\sim P}\big[h(x) \neq y\big].$$

We can only observe the *training error* $\mathrm{TrainErr}_n(h)$. Finite-sample guarantees take the form

$$\Big|\mathrm{TrainErr}_n(h) - \mathrm{TrueErr}(h)\Big| \;\leq\; \sqrt{\frac{c(\mathcal{H})}{n}}$$

with probability at least $1-\delta$, where $c(\mathcal{H})$ is a complexity parameter (VC dimension, $\log|\mathcal{H}|$, Rademacher, etc.) [1].

Homework link. Exercise 2 asks which photo-collection strategy respects this i.i.d. assumption. Training on images sampled *across California* best matches the test distribution, hence minimises hidden covariate shift.

2.1 Margin-based Bound

For linear separators that classify all points with margin $\gamma > 0$ and radius R $\big(\|x^{(i)}\| \leq R\big)$, one obtains

$$c(\mathcal{H}) \;\leq\; \frac{R^2}{\gamma^2} \quad \text{(independent of } d\text{)},$$

explaining why large-margin SVMs can succeed in extreme dimension-to-sample-ratio settings [2].

Homework 7, Problem 3. Even with one million features and only 1000 points, a large margin keeps R^2/γ^2 modest, so the generalisation gap shrinks.

3 Multiclass Linear Predictors

For each class j store parameters $(w_j, b_j) \in \mathbb{R}^{d+1}$. Define the score function

$$f_j(x) = w_j^\top x + b_j, \qquad \hat{y} = \arg\max_j f_j(x).$$

Decision regions are convex polytopes separated by $(w_j - w_\ell)^\top x + (b_j - b_\ell) = 0$ [3].

3.1 Worked Boundary Example (Homework 1)

Given $w_1 = (1,1)$, $b_1 = 0$, $w_2 = (1,0)$, $b_2 = 1$, $w_3 = (0,1)$, $b_3 = -1$, the pairwise hyperplanes are $y = 1$, $x = -1$, and $y = x + 2$. Figure 1 diagrams the regions.

3.2 Multiclass Perceptron

Loss Function

$$\ell_{\text{perc}}((x,y); W) = \begin{cases} 0 & f_y(x) \geq f_j(x)\ \forall j, \\ 1 & \text{otherwise.} \end{cases}$$

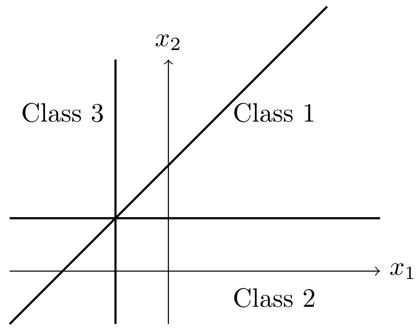

Figure 1: Decision regions for the boundary example.

Algorithm

Multiclass Perceptron (one pass)
input data $(x^{(i)}, y^{(i)})_{i=1}^{n}$, classes $1{:}k$
Initialise $w_j \leftarrow 0,\ b_j \leftarrow 0$
for $i = 1$ to n:
  predict $\hat{y} \leftarrow \arg\max_j (w_j^\top x^{(i)} + b_j)$
  if $\hat{y} \neq y^{(i)}$:
    $w_{y^{(i)}} \leftarrow w_{y^{(i)}} + x^{(i)}, \quad w_{\hat{y}} \leftarrow w_{\hat{y}} - x^{(i)}$
    $b_{y^{(i)}} \leftarrow b_{y^{(i)}} + 1, \quad b_{\hat{y}} \leftarrow b_{\hat{y}} - 1$

Convergence Theorem If there exist parameters with margin $\gamma > 0$ on data $\|x^{(i)}\| \leq R$, the total number of mistakes is no more than $(R/\gamma)^2$ [4].

3.3 Soft-Margin SVM (Crammer–Singer)

Primal Problem

$$\min_{w_j, b_j, \xi} \sum_{j=1}^{k} \|w_j\|_2^2 + C \sum_{i=1}^{n} \xi_i \tag{1}$$
$$\text{s.t. } f_{y^{(i)}}(x^{(i)}) - f_y(x^{(i)}) \geq 1 - \xi_i, \quad \forall i,\ \forall y \neq y^{(i)},\ \xi_i \geq 0.$$

Variables $= kd + k + n$; constraints $= n(k-1)$.

Margin/Complexity Trade-off Large C penalises slack heavily, yielding wider margins but potentially higher variance; small C allows violations, trading bias for variance.

3.4 Multinomial Logistic Regression

Softmax Model

$$p(y = j \mid x) = \frac{\exp(f_j(x))}{\sum_{\ell=1}^{k} \exp(f_\ell(x))}, \qquad \hat{y} = \arg\max_j f_j(x).$$

Objective Minimise $-\sum_{i=1}^{n} \log p(y^{(i)} \mid x^{(i)})$, a convex function in (w_j, b_j).

Calibration Reliability diagrams check whether predicted probabilities match empirical frequencies; temperature scaling can correct mis-calibration.

4 Implementation Details

Quick Python Snippets

```
def mc_perceptron(X, y, k, epochs=30):
    d = X.shape[1]
    W = np.zeros((k, d))
    b = np.zeros(k)
    for _ in range(epochs):
        for xi, yi in zip(X, y):
            scores = W.dot(xi) + b
            y_hat = scores.argmax()
            if y_hat != yi:
                W[yi] += xi
                W[y_hat] -= xi
                b[yi] += 1
                b[y_hat] -= 1
    return W, b
```

```
from sklearn.svm import LinearSVC
clf = LinearSVC(loss="hinge", multi_class="crammer_singer", C=1.0)
clf.fit(X_train, y_train)
```

5 Distribution Shift

Covariate Shift $P_{tr}(x) \neq P_{te}(x)$, but $P(y \mid x)$ fixed. *Importance weighting*: weight test loss by $w(x) = \frac{P_{te}(x)}{P_{tr}(x)}$.

Label Shift $P_{tr}(y) \neq P_{te}(y)$, and $P(x \mid y)$ unchanged. Estimate new class priors by moment-matching $\hat{\pi} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \hat{q}$ where $\mathbf{X}_{ij} = p(x^{(i)} \mid y = j)$.

Homework 7, Problem 4 answers. (a) Different topic priors $\Rightarrow$ label shift. (b) Same topics but vocabulary drift $\Rightarrow$ covariate shift.

6 Mathematical Justification

6.1 Why Margin Helps Generalisation

The margin bound R^2/γ^2 arises from bounding the growth function of half-spaces separated by margin [2]. Maximising the minimum margin (as SVM does) directly shrinks the complexity term, lowering required sample size.

6.2 Why Perceptron Converges (Sketch)

Let $u = [w_1^*; \ldots; w_k^*]$ be a reference separator with unit norm and margin γ. Define global parameter vector θ by stacking all w_j. Each mistake raises $u^\top \theta$ by at least γ and increases $\|\theta\|$ by at most R. After M mistakes, $\gamma M \leq u^\top \theta \leq \|\theta\| \leq R\sqrt{M}$, hence $M \leq (R/\gamma)^2$.

7 Programming-Exercise Checklist

1. Draw decision regions on a 400×400 grid using `plt.contourf`.
2. Shuffle data after each Perceptron epoch; early-stop if no errors.
3. Train SVM for $C = 0.01, 0.1, 1, 10$ and plot four separate boundaries. Comment on margin width versus boundary jaggedness.

8 Conclusion

Perceptron, SVM, and logistic regression share a common geometric core; their differences lie in loss functions and regularisers. Margin maximisation links simplicity to generalisation, and understanding covariate versus label shift is critical for reliable deployment.

References

[1] Vladimir Vapnik. *Statistical Learning Theory.* Wiley, 1998.

[2] Peter Bartlett and Philip Bartlett. The Sample Complexity of Pattern Classification with Margin. In *IEEE Transactions on Information Theory*, 1998.

[3] Koby Crammer and Yoram Singer. On the Algorithmic Implementation of Multiclass Kernel-Based Vector Machines. *Journal of Machine Learning Research*, 2001.

[4] Koby Crammer and Yoram Singer. Ultraconservative Online Algorithms for Multiclass Problems. *Journal of Machine Learning Research*, 2003.