Multiclass linear prediction

DSE 220

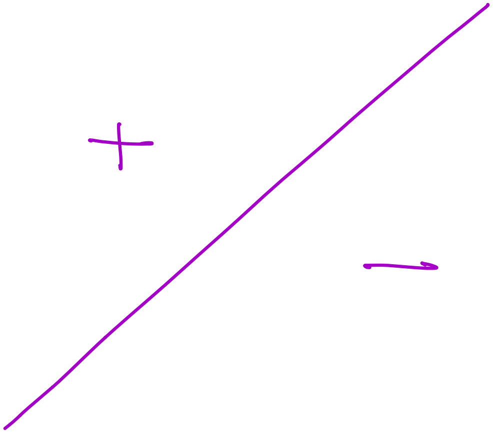
Topics we'll cover

1. Multiclass logistic regression
2. Multiclass Perceptron
3. Multiclass support vector machines

Multiclass classification

Of the classification methods we have studied so far, which seem inherently binary?

- Nearest neighbor?
- Generative models?
- Linear classifiers?



The main idea

Remember Gaussian generative models...

Each class j has an associated function

$$F_j(x) = \ln\Big(\underbrace{\pi_j}_{\text{weight of class } j}\ \underbrace{P_j(x)}_{\text{Gaussian density for class } j}\Big)$$

Gaussian case:

- $F_j(x)$ quadratic
- But linear if covariance matrices all equal

To classify a new point x:

- Evaluate $F_1(x), F_2(x), \ldots, F_k(x)$ [if k classes]
- The biggest value wins → we predict that class

For linear classification, each class $(j = 1, 2, .., k)$ gets its own linear function.

Class 1: $w_1 \cdot x + b_1$

Class 2: $w_2 \cdot x + b_2$

$\vdots$

Class k: $w_k \cdot x + b_k$

if $x \in \mathbb{R}^d$ then $w_1, .., w_k \in \mathbb{R}^d$ and $b_1, .., b_k \in \mathbb{R}$

$\boxed{\text{Linear classification in the \underline{multiclass} setting}}$

To predict the class of a new point x:

- Evaluate all k of these functions
- Largest value wins

From binary to multiclass logistic regression

Binary logistic regression: for $\mathcal{X} = \mathbb{R}^d$, classifier given by $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$:

$$\Pr(y = 1|x) = \frac{e^{w \cdot x + b}}{1 + e^{w \cdot x + b}}$$

From binary to multiclass logistic regression

Binary logistic regression: for $\mathcal{X} = \mathbb{R}^d$, classifier given by $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$:

$$\Pr(y = 1|x) = \frac{e^{w \cdot x + b}}{1 + e^{w \cdot x + b}}$$

Labels $\mathcal{Y} = \{1, 2, \ldots, k\}$: specify a classifier by $w_1, \ldots, w_k \in \mathbb{R}^d$ and $b_1, \ldots, b_k \in \mathbb{R}$:

$$\Pr(y = j|x) \;\propto\; e^{w_j \cdot x + b_j}$$

From binary to multiclass logistic regression

Binary logistic regression: for $\mathcal{X} = \mathbb{R}^d$, classifier given by $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$:

$$\Pr(y = 1|x) = \frac{e^{w \cdot x + b}}{1 + e^{w \cdot x + b}}$$

Labels $\mathcal{Y} = \{1, 2, \ldots, k\}$: specify a classifier by $w_1, \ldots, w_k \in \mathbb{R}^d$ and $b_1, \ldots, b_k \in \mathbb{R}$:

$$\Pr(y = j|x) \underset{\text{proportional}}{\propto} e^{w_j \cdot x + b_j}$$

- What is the fully normalized form of the probability?

$$\Pr(y=j|x) = \frac{e^{w_j \cdot x + b_j}}{\sum_{\ell=1}^{k} e^{w_\ell \cdot x + b_\ell}}$$

"softmax"

- Given a point x, which label to predict?

$$\arg\max_j \; w_j \cdot x + b_j$$

Multiclass logistic regression

- **Label space**: $\mathcal{Y} = \{1, 2, \ldots, k\}$
- **Parametrized classifier**: $w_1, \ldots, w_k \in \mathbb{R}^d$, $b_1, \ldots, b_k \in \mathbb{R}$:

$$\Pr(y = j | x) = \frac{e^{w_j \cdot x + b_j}}{e^{w_1 \cdot x + b_1} + \cdots + e^{w_k \cdot x + b_k}}$$

- **Prediction**: given a point x, predict label $\arg\max_j \, (w_j \cdot x + b_j)$.
- **Learning**: Given: $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$. $x^{(i)} \in \mathbb{R}^d$, $y^{(i)} \in \{1, 2, .., k\}$
  Find: $w_1, \ldots, w_k \in \mathbb{R}^d$ and $b_1, \ldots, b_k$ that maximize the likelihood

$$\prod_{i=1}^{n} \Pr(y^{(i)} | x^{(i)})$$

Taking negative log gives a convex minimization problem.

Multiclass Perceptron

Setting: $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \{1, 2, \ldots, k\}$

Model: $w_1, \ldots, w_k \in \mathbb{R}^d$ and $b_1, \ldots, b_k \in \mathbb{R}$

Prediction: On instance x, predict label $\arg\max_j (w_j \cdot x + b_j)$

Multiclass Perceptron

Setting: $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \{1, 2, \ldots, k\}$

Model: $w_1, \ldots, w_k \in \mathbb{R}^d$ and $b_1, \ldots, b_k \in \mathbb{R}$

Prediction: On instance x, predict label $\arg\max_j (w_j \cdot x + b_j)$

Learning. Given training set $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$:

- Initialize $w_1 = \cdots = w_k = 0$ and $b_1 = \cdots = b_k = 0$
- Repeat while some training point (x, y) is misclassified:

only update two classes

for correct label y: $\quad w_y = w_y + x$

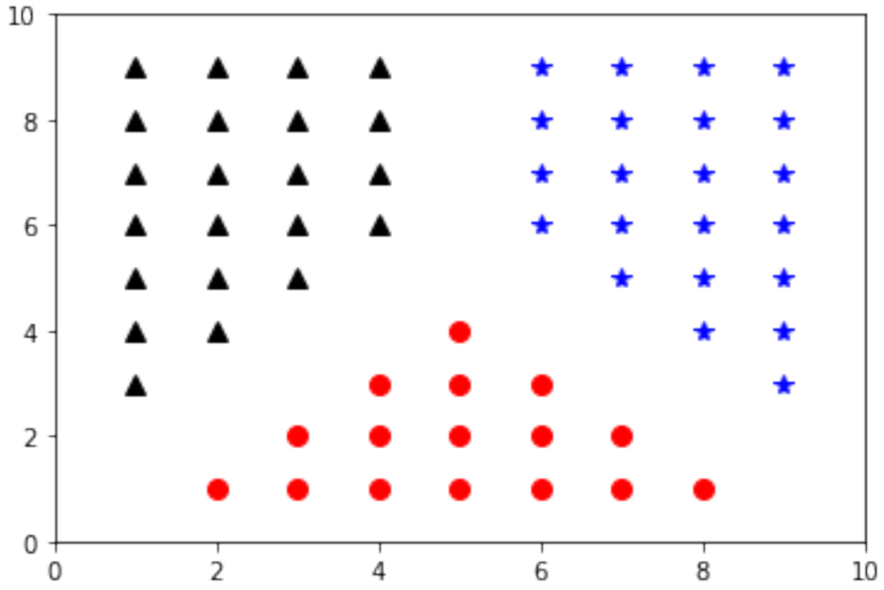
$\qquad\qquad\qquad\qquad b_y = b_y + 1$

for predicted label $\widehat{y}$: $\quad w_{\widehat{y}} = w_{\widehat{y}} - x$

$\qquad\qquad\qquad\qquad b_{\widehat{y}} = b_{\widehat{y}} - 1$

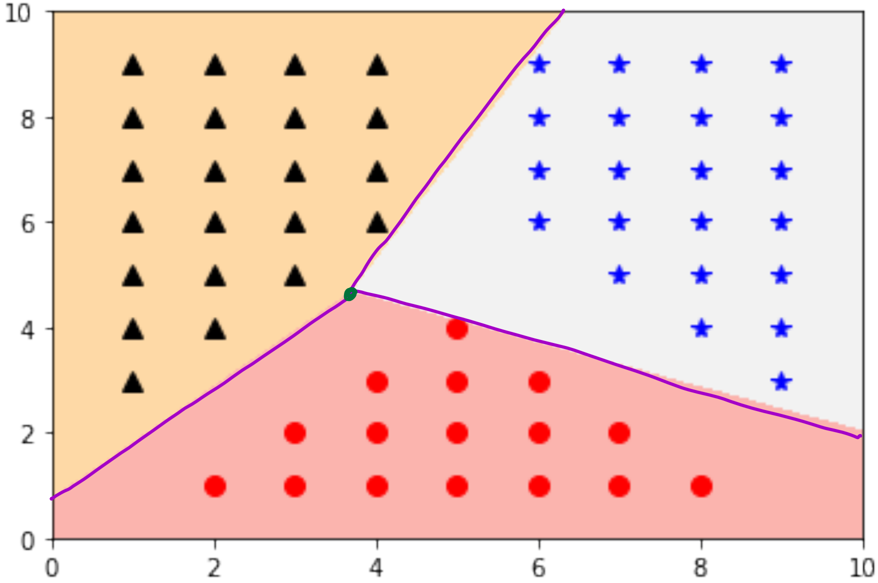
$\hat{y}$ is $\arg\max_j w_j \cdot x + b_j$

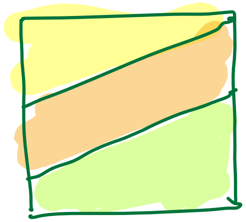
Multiclass Perceptron: example



Multiclass Perceptron: example

decision boundary is composed of linear pieces





Multiclass SVM

Model: $w_1, \ldots, w_k \in \mathbb{R}^d$ and $b_1, \ldots, b_k \in \mathbb{R}$

Prediction: On instance x, predict label $\arg\max_j (w_j \cdot x + b_j)$

Learning. Given training set $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$:

maximize margins

penalty for using slack

Convex program

$$\min_{w_1,\ldots,w_k \in \mathbb{R}^d, b_1,\ldots,b_k \in \mathbb{R}, \xi \in \mathbb{R}^n} \sum_{j=1}^{k} \|w_j\|^2 + C \sum_{i=1}^{n} \xi_i$$

$$w_{y^{(i)}} \cdot x^{(i)} + b_{y^{(i)}} - w_y \cdot x^{(i)} - b_y \geq 1 - \xi_i \quad \text{for all } i, \text{ all } y \neq y^{(i)}$$?

$$\xi \geq 0$$

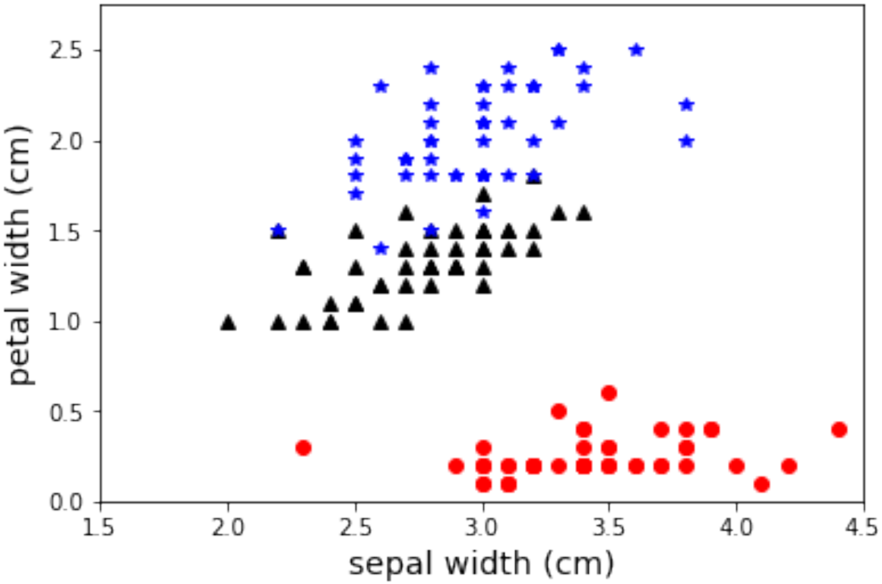
Point $x^{(i)}$ has true label $y^{(i)}$

$\therefore$ Want $W_{y^{(i)}} \cdot x^{(i)} + b_{y^{(i)}} > W_y \cdot x^{(i)} + b_y$ for all other labels $y \neq y^{(i)}$

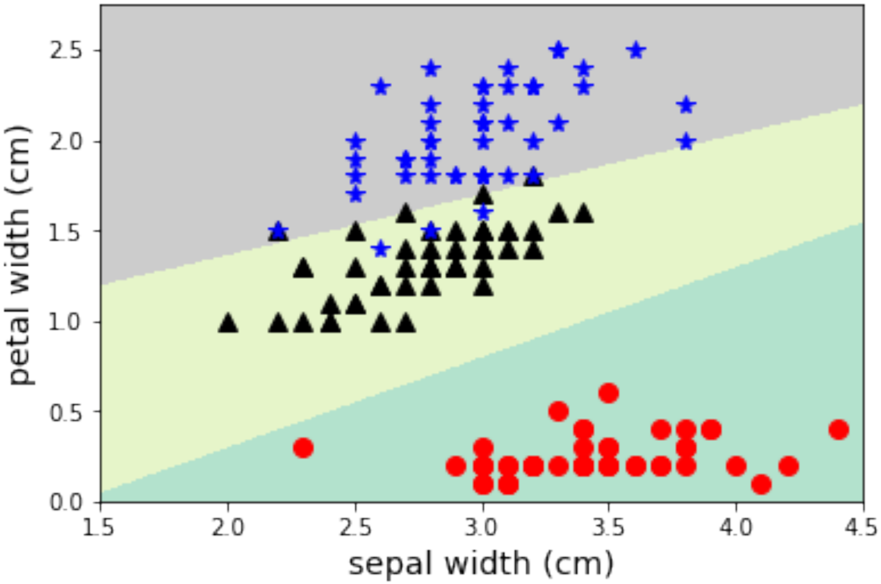
As before, replace with $W_{y^{(i)}} \cdot x^{(i)} + b_{y^{(i)}} \geq W_y \cdot x^{(i)} + b_y + 1 - \xi_i$

slack

Multiclass SVM example: iris



Multiclass SVM example: iris



Multiclass SVM

Given training set $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$:

$$\boxed{\begin{gathered}\min_{w_1,\ldots,w_k \in \mathbb{R}^d, b_1,\ldots,b_k \in \mathbb{R}, \xi \in \mathbb{R}^n} \sum_{j=1}^{k} \|w_j\|^2 + C \sum_{i=1}^{n} \xi_i \\ w_{y^{(i)}} \cdot x^{(i)} + b_{y^{(i)}} - w_y \cdot x^{(i)} - b_y \geq 1 - \xi_i \quad \text{for all } i, \text{ all } y \neq y^{(i)} \\ \xi \geq 0 \end{gathered}}$$

Once again, a convex optimization problem.

Question: how many variables and constraints do we have?

<u>Variables</u>

① The classifiers: $k(d+1)$

② Slack variables: n

<u>Constraints</u>

For each data pt: want the correct label to beat the remaining $k-1$ labels.

$\therefore$ Total of $n(k-1)$ constraints.

Back to binary setting:

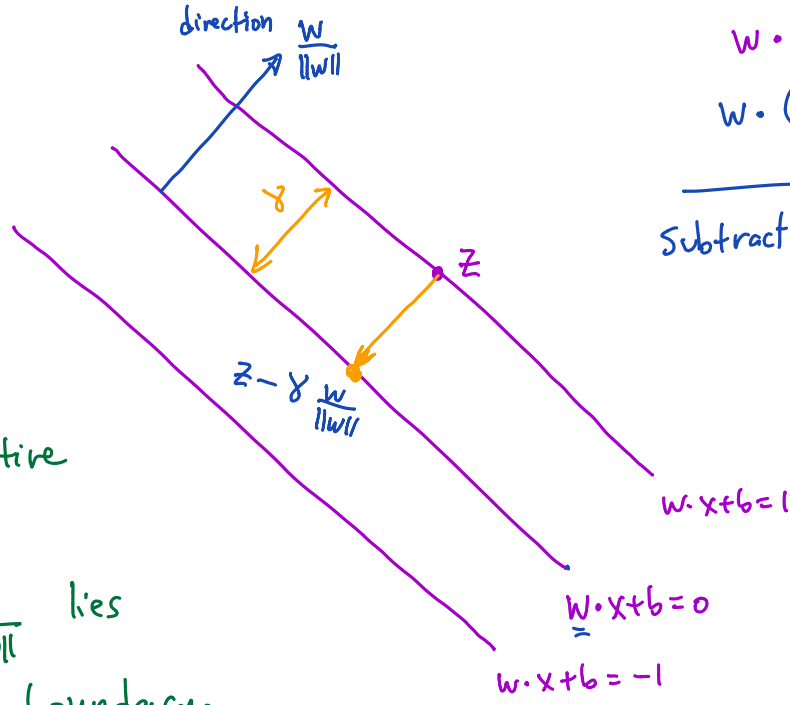
$$y^{(i)}\,(w\cdot x^{(i)} + b) > 0 \qquad \text{for all } i = 1..n$$

$\Uparrow \quad \Downarrow$ multiply w, b by a large enough constant

$$y^{(i)}\,(w\cdot x^{(i)} + b) \geq 1 \qquad \text{for all } i = 1..n$$

if in this form, have a nice expression for the margin

(margin $= 1/\|w\|$)



$$w \cdot z + b = 1$$

$$w \cdot \left(z - \gamma \frac{w}{\|w\|}\right) + b = 0$$

Subtract: $\gamma \|w\| = 1$

$$\Rightarrow \gamma = \frac{1}{\|w\|}$$

① Let z be <u>any</u> point on the positive boundary.

② Then $z - \gamma \frac{w}{\|w\|}$ lies on the decision boundary.

Worksheet 10

#1