



## تقرین سوم

### بارمبندی

بارم سوالات بدین شکل است:

- سوال ۱: ۲۵ امتیاز
- سوال ۲: ۲۰ امتیاز
- سوال ۳: ۲۰ امتیاز
- سوال ۴: ۳۰ امتیاز
- سوال ۵: ۱۵ امتیاز

### مسئله ۱. (صفحه جداکننده)

- فرض کنید  $X$  و  $Y$  دو مجموعه متناهی از نقاط در  $\mathbb{R}^n$  هستند؛ به طوری که ترکیب خطی آفین<sup>۱</sup> دو مجموعه با یکدیگر اشتراک ندارند. ثابت کنید صفحه‌ای جداکننده برای نقاط مجموعه‌های  $X$  و  $Y$  وجود دارد.
- فرض کنید که  $n + 1$  نقطه در  $[0, 1]^n$  به صورت یکنواخت انتخاب کرده‌ایم و به هر یک از آنها رنگ قرمز یا آبی نسبت داده ایم. ثابت کنید با احتمال بیشتر از نصف نقاط دو رنگ قابل جداسازی‌اند.
- در مسئله‌ی SVC، حداقل تعداد نقاط بردارهای پشتیبان<sup>۲</sup> را بر حسب  $n$  تعیین کنید.

### مسئله ۲. (درخت تصمیم گیری)

- معیارهای Entropy Information، Gini Impurity و Gain را در درخت های تصمیم به طور مختصر توضیح دهید.

<sup>۱</sup> ترکیب خطی آفین مجموعه  $X$  را بردارهایی در نظر می‌گیریم که به شکل  $\sum_{x \in X} \alpha_x x$  قابل نمایش هستند؛ به طوری که  $\sum \alpha_x = 1$  و  $\alpha_x \geq 0$ .

<sup>۲</sup>Support Vectors

- معیارهای Gini و Entropy را از نظر بازه جوابشان مقایسه کنید. با توجه به تعریف ریاضیاتی هر کدام، به نظرتان استفاده از هریک چه مزایا و معایبی نسبت به دیگری دارد؟
- روی داده‌های زیر با متغیرهای  $x_1$  و  $x_2$  و متغیر هدف  $y$  یک درخت تصمیم با معیار gini آموزش دهید و سپس مقدار  $y$  را برای داده‌ی جدید ( $x_1 = 1/5, x_2 = 1/5$ ) پیش‌بینی کنید. آیا با انتخاب معیار entropy حاصل تغییر می‌کند؟

$x_1$	$x_2$	$y$
1.5	2.0	1
3.0	2.8	0
1.2	0.7	1
4.2	2.1	1
1.5	1.8	0
0.9	1.2	0
3.1	2.1	1
2.7	2.3	1
1.0	0.8	0
2.8	1.9	1

### مسئله‌ی ۳. (Spline)

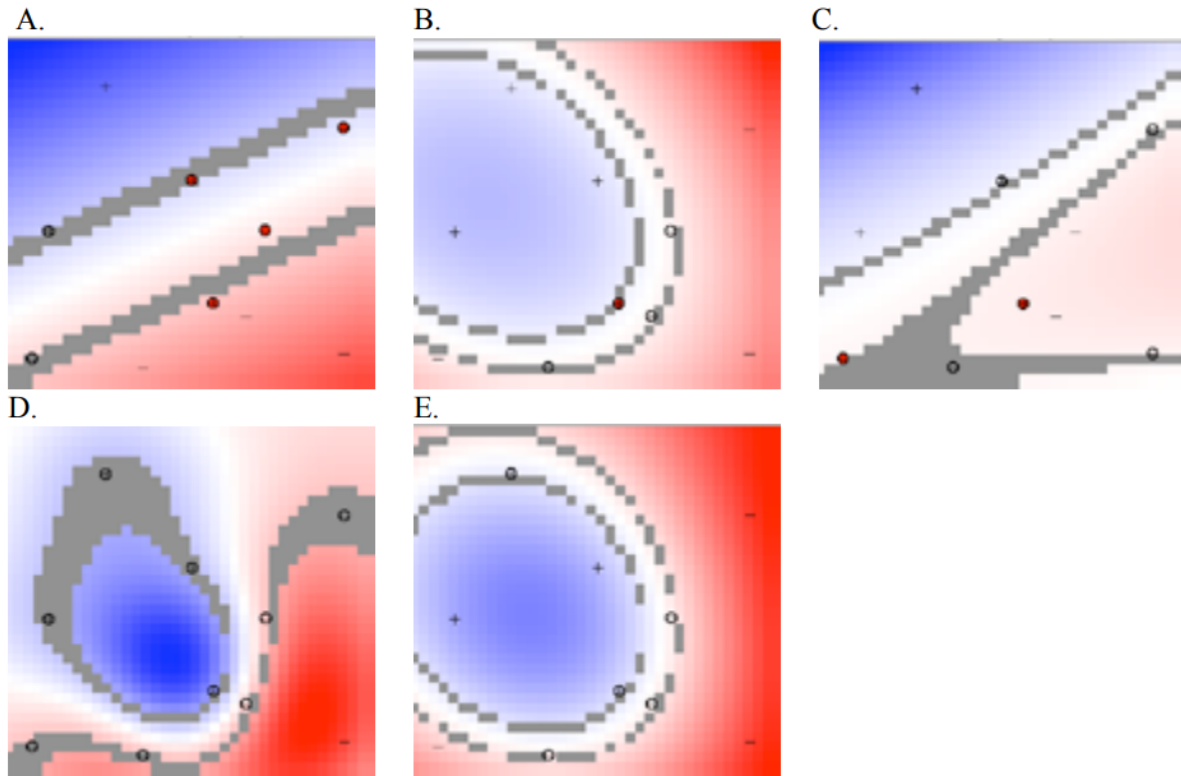
- فرض کنید که نقاط  $(x_i, y_i) \in \mathbb{R}^2$  به ازای  $i \in \{1, 2, \dots, n\}$  داده شده است. تعداد گره‌ها  $k$  را در نظر بگیرید. درجه آزادی natural spline و cubic spline را پیدا کنید.
- فرض کنید علاوه بر کمینه کردن مربعات، تابع  $\lambda \int f''(x)^2 dx$  را به تابع هزینه <sup>۴</sup> اضافه کنیم. اگر  $\lambda$  به سمت بینهایت برود، تابعی که توسط Natural Spline بدست می‌آید را تحلیل کنید.
- فرض کنید که می‌خواهید spline را برای ابعاد بالاتر در نظر بگیرید، مدلی ارائه دهید و در آن گره‌ها را تعریف کنید و همچنین درجه‌ی آزادی مدل خود را بررسی کنید. معایب مدل خود را بیان کنید.

---

<sup>۳</sup> knot  
<sup>۴</sup> function Loss

#### مسئله ۴. (ماشین پشتیبان بردار)

- نمودارهای زیر مربوط به SVM هایی هستند که وظیفه جداسازی + ها را از - ها برعهده دارند. لازم به ذکر است که این مدل ها بر روی یک مجموعه داده یکسان آموزش دیده اند. کدام یک از این نمودارها نمایانگر بهترین دسته بند برای داده های آموزشی است؟



- هر کدام از نمودارهای بالا را به یکی از کرنل های زیر اختصاص دهید و راجع به نحوه تصمیم گیری خود مختصر توضیح دهید.

1. Radial basis function, sigma .08
2. Radial basis function, sigma .5
3. Radial basis function, sigma 2.0
4. linear
5. second order polynomial

- یک SVM دسته بند را برای جداسازی مثبت ها از منفی ها در نظر بگیرید. فرض کنید این ماشین یک بردار پشتیبان مثبت در نقطه  $x_1 = (1, 0)$  و یک بردار پشتیبان منفی در نقطه  $x_2 = (0, 1)$  دارد. مقدار  $w$  و  $b$  را برحسب  $x_1$  و  $x_2$  بدست آورید.

## مسئله‌ی ۵. ... Bagging and

- در Boosting بررسی کنید که چه پارامترهایی باید برای یادگیری تعیین شوند؟ کدام مورد باعث overfit می‌شود ولی در Bagging منجر به این پدیده نمی‌شود؟ توضیح دهید.
- یک مدل درخت تصمیم‌گیری ساده را با مدل *Random Forest* مقایسه کنید (یک عامل برتری و ضعف درخت ساده را بیان کنید).