Aprendizagem de Máquina I – Avaliação Presencial 02

Prof. Hugo Carvalho

02/12/2025

Questão 1: Considere o contexto de classificação binária, ou seja, $\mathbf{X} \in \mathbb{R}^p$ é um vetor aleatório de dimensão p ao qual atribuímos uma classe $Y \in \mathcal{C} = \{0, 1\}$. Assuma que a distribuição conjunta do par aleatório $(\mathbf{X}, Y)$ é conhecida, e por simplicidade assuma também que o vetor aleatório $\mathbf{X}$ é contínuo. O problema de classificação consiste em encontrar um classificador $g : \mathbb{R}^p \to \mathcal{C}$, com base em um conjunto de observações, que satisfaça algum critério de interesse. Para a presente questão, considere que tal critério é minimizar a função risco a seguir:

$$\begin{aligned} R(g) &= \mathbb{E}[\ell_1 \mathbb{I}(Y \neq g(\mathbf{X}) \text{ e } Y = 0) + \ell_0 \mathbb{I}(Y \neq g(\mathbf{X}) \text{ e } Y = 1)] \\ &= \ell_1 \mathbb{P}(Y \neq g(\mathbf{X}) \text{ e } Y = 0) + \ell_0 \mathbb{P}(Y \neq g(\mathbf{X}) \text{ e } Y = 1). \end{aligned}$$

Com base nisso, faça o que se pede abaixo.

a) Mostre que a função g que minimiza tal risco é dada por:

$$g(\mathbf{x}) = \begin{cases} 1, & \text{se } \mathbb{P}(Y = 1 | \mathbf{X} = \mathbf{x}) > \frac{\ell_1}{\ell_0 + \ell_1}, \\ 0, & \text{caso contrário.} \end{cases}$$

b) Atribua os rótulos "positivo" e "negativo" às classes 1 e 0, respectivamente. Suponha que cometer um falso positivo seja 5 vezes pior que cometer um falso negativo. Explicite e interprete a regra de decisão encontrada pelo resultado acima nesse cenário.

Questão 2: A tabela abaixo ilustra um conjunto de $n = 7$ observações em $p = 2$ dimensões, juntamente com as suas respectivas classes:

No. da obs.	X_1	X_2	Classe
1	3	4	Vermelho
2	2	2	Vermelho
3	4	4	Vermelho
4	1	4	Vermelho
5	2	1	Azul
6	4	3	Azul
7	4	1	Azul

Com base nisso, faça o que se pede abaixo:

a) Finja que você está ajustando um SVM com *kernel* linear para esses dados. Esboce o que parece ser a reta que separa esse conjunto de dados, considerando a formulação teórica do SVM. Ilustre também a margem dessa reta e os vetores suporte. Justifique detalhadamente suas escolhas.

b) Argumente que uma pequena modificação na observação de número 7 não mudará o plano que você esboçou.

c) Suponha que você deseja classificar uma nova observação, de coordenadas $(3, 3)$, a partir desses dados, agora usando o KNN. Qual seria a classe atribuída usando $K = 1$ e $K = 3$?

Questão 3: Explique detalhadamente o algoritmo para se ajustar uma árvore de decisão em um problema de classificação. Para simplificar, assuma que o problema é de classificação binária, ou seja, $\mathcal{C} = \{0, 1\}$. Discuta também as métricas utilizadas para medir a qualidade de um novo ramo na árvore, análogas ao EQM no contexto de regressão:

- Erro de classificação: $E = 1 - \max_c p_{R,c}$
- Índice de Gini: $G = \sum_{c \in \mathcal{C}} p_{R,c}(1 - p_{R,c})$
- Entropia: $D = -\sum_{c \in \mathcal{C}} p_{R,c} \log(p_{R,c})$.

Lembre-se que $p_{R,c}$ é a proporção de observações classificadas como sendo da classe c entre as que caem na região R.

① (a) Note que minimizar $R(g)$ é equivalente a minimizar $R(g(x))$, para todo $x \in \mathbb{R}^d$. Calculemos quem é tal quantidade:

$$R(g) = l_1 \mathbb{P}(Y \neq g(X) \text{ e } Y=0) + l_0 \mathbb{P}(Y \neq g(X) \text{ e } Y=1)$$

$$= l_1 \mathbb{P}(Y=0 \mid X=x)\mathbb{P}(g(X)=1) + l_0 \mathbb{P}(Y=1 \mid X=x)\mathbb{P}(g(X)=0)$$

$$\Rightarrow R(g(x)) = \begin{cases} l_0 \mathbb{P}(Y=1 \mid X=x), & \text{se } g(x)=0; \\ l_1 \mathbb{P}(Y=0 \mid X=x), & \text{se } g(x)=1. \end{cases}$$

Portanto, para minimizar $R(g(x))$, escolheremos:

$$g(x) = \begin{cases} 1, & \text{se } \underbrace{l_1 \mathbb{P}(Y=0 \mid X=x) < l_0 \mathbb{P}(Y=1 \mid X=x)}_{\star} \\ 0, & \text{caso contrário.} \end{cases}$$

Manipulando a desigualdade $\star$, temos que:

$$l_1 \mathbb{P}(Y=0 \mid X=x) < l_0 \mathbb{P}(Y=1 \mid X=x)$$

$$\Leftrightarrow l_1 [1 - \mathbb{P}(Y=1 \mid X=x)] < l_0 \mathbb{P}(Y=1 \mid X=x)$$

$$\Leftrightarrow \mathbb{P}(Y=1 \mid X=x) > \frac{l_1}{l_0 + l_1},$$

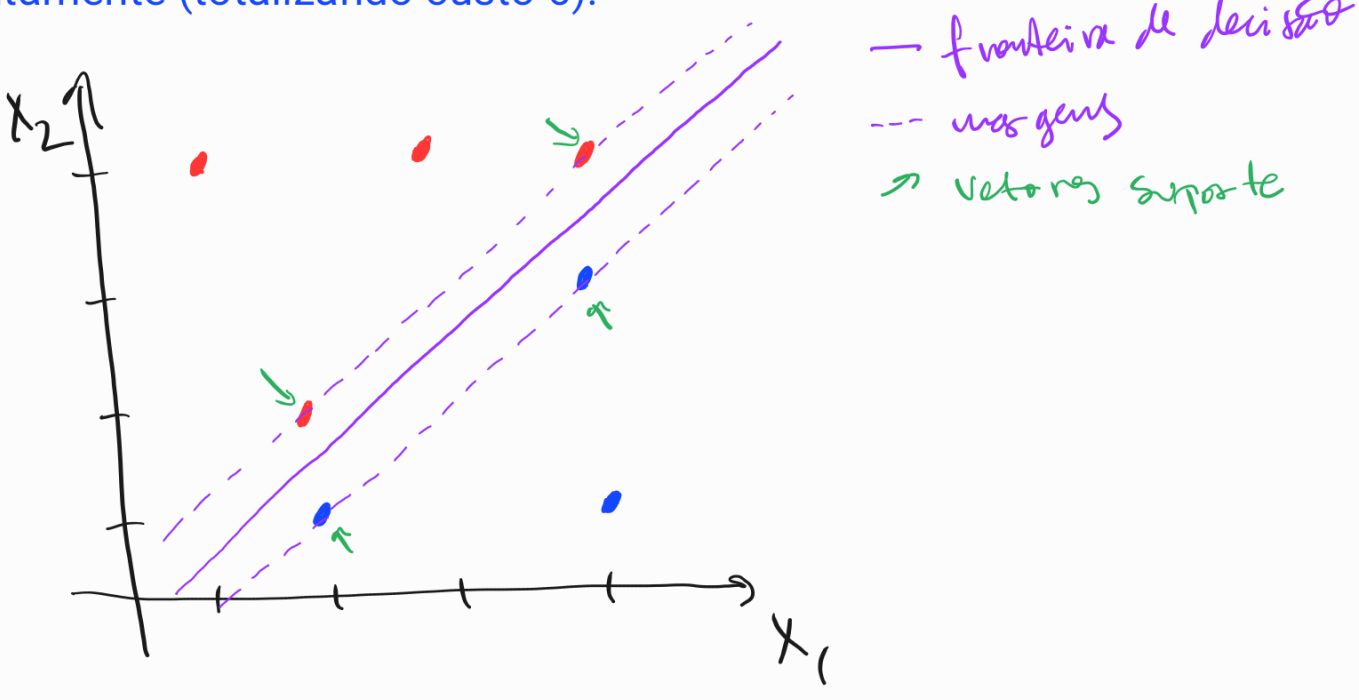
conforme o enunciado.

b) Se cometer um falso positivo é 5 vezes pior que cometer um falso negativo, temos que os pesos são dados por: $\ell_1 = 5$ e $\ell_0 = 1$.
Dessa forma, a regra de decisão é dada por:

$$g(x) = \begin{cases} 1, \text{ se } \mathbb{P}(Y=1 \mid X=x) > \frac{5}{6}, \\ 0, \text{ caso contrário.} \end{cases}$$

Ou seja, dado esse risco, a fim de evitar falsos positivos, torna-se mais "difícil" classificar uma observação como positiva. O valor de $5/6$ é dado pelo teorema provado no item a), a fim de minimizar o risco de classificação. Ele pode ser interpretado como um "peso relativo" de um tipo de erro em relação (de custo 5) aos dois tipos de erros conjuntamente (totalizando custo 6).

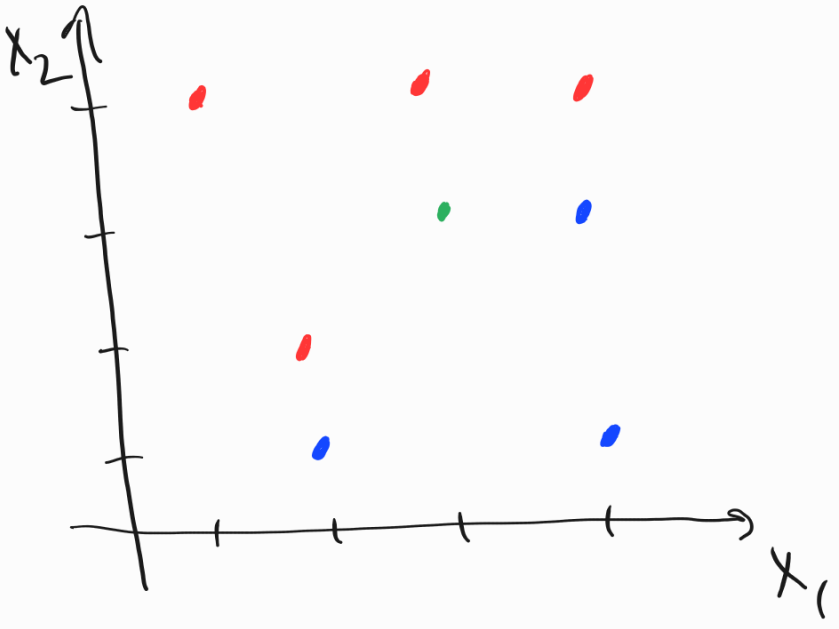
(2) (a)



A fronteira de decisão linear dada pelo SVM com *kernel* liner é uma reta que irá separar pefeitamente as duas classes, já que o problema é linearmente separável, e estará "igualmente longe" das observações mais próximas da fronteira de decisão. Isso justifica a reta desenhada bem como a sua margem. Os vetores suporte são os vetores que definem a margem, ou seja, as observações mais próximas à fronteira de decisão, que nesse caso têm coordenadas (2,1), (2,2), (4,3) e (4,4).

b) A observção de índice 7 tem coordenadas (4,1), estando, portanto, bastante longe da margem. Qualquer modificação dela que não a faça "bater"na margem não irá mudar a fronteira de decisão desse classificador, e nem da sua margem. Lembremos que em um SVM, independente do *kernel* utilizado, apenas os vetores suporte são utilizados para construir a fronteira de decisão, sendo os vetores mais afastados de "menor importância".

c)



Note que a nova observação está igualmente próxima das observações de coordenadas (3,4) (da classe vermelha) e (4,3)(da classe azul). Dessa forma, com K = 1, é impossível tomar uma decisão para essa observação, devido a esse "empate". Note que essas também são as duas observações mais próximas, de modo que com K = 2 haveria um "empate" na votação das classes. Agora, com K = 3, acrescentamos a essas observações a de coordenadas (4,4) ou (2,2), já que ambas têm a mesma distância para (3,3), e felizmente ambas são da classe vermelha. Nesse contexto, dentre as três observações mais próximas, temos duas vermelhas e uma azul, de modo que a classificação será pela classe vermelha.

3) Para ajustar uma árvore de decisão no contexto de classificação binária, assim como em qualquer outro caso, divide-se o conjunto de dados em treinamento e teste. No conjunto de treinamento realizamos o seguinte procedimento iterativo: busca-se pelo "melhor" atributo para levar em consideração para a criação de um novo nó, e para esse

"melhor" atributo escolhe-se o "melhor" ponto de corte para definir esse nó. Exemplificando, em determinado passo da criação da árvore pode se descobrir que o "melhor" atributo é o X_5, e o "melhor" ponto de corte é 7, de modo que o novo nó da árvore irá criar dois novos ramos, a saber, X_5 < 7 e X_5 > 7, implicando em decisões diferentes para cada uma dessas possibilidades. O critério que se usa para encontrar esses "melhores" valores diz respeito à "pureza" do novo nó criado, ou seja, deseja-se que um novo nó seja o mais uniforme possível, com a maior concentração de observações da mesma classe em um ramo ou outro do nó. No exemplo acima, imagina-se que as regiões X_5 < 7 e X_5 > 7 (levando em conta as restriões já impostas pelos nós anteriores da árvore, claro) tenham o máximo de concentração de observações da mesma classe. Esse processo iterativo é feito até que algum criério seja satisfeito. Em aula usualmente adotamos o critério de cada nó terminal ter pelo menos 5 observações, a fim de evitar o sobreajuste.

As três funções informadas são utilizadas para medir a "pureza" de um novo nó a ser criado, e o fazem de maneira diferente. No contexto de classificação binária, vamos imaginar que temos $p_{R,1} \cong 1$, e portanto, $p_{R,0} \cong 0$. Teremos então que:

$$E = 1 - \max_c p_{R,c} = 1 - p_{R,1} \cong 0$$

$$G = \sum_{c \in \{0,1\}} p_{R,c}(1 - p_{R,c}) = \underbrace{p_{R,0}}_{\cong 0}\underbrace{(1 - p_{R,0})}_{\cong 1} + \underbrace{p_{R,1}}_{\cong 1}\underbrace{(1 - p_{R,1})}_{\cong 0} \cong 0$$

$$D = -\sum_{c \in \{0,1\}} p_{R,c} \log(p_{R,c}) = -\Big[\underbrace{p_{R,0}}_{\cong 0}\underbrace{\log(p_{R,0})}_{\cong -\infty} + \underbrace{p_{R,1}}_{\cong 1}\underbrace{\log(p_{R,1})}_{\cong 0}\Big] \cong 0$$

$\cong 0$, por L'Hôpital

Ou seja, essas três quantidades são pequenas quando um novo nó é bastante puro. A interpretação delas é a seguinte:

- O erro de classificação parte do pressuposto que a classe majoritária em uma região é a "certa", de modo que todas as outras estão "erradas", sendo essa proporção "errada"capturada pela

quantidade.

- A entropia, conforme discutida em sala de aula, mede a "surpresa média" de uma variável aleatória. Assim, uma baixa entropia significa uma variável aleatória com "baixa surpresa", ou seja, há um evento muito provável e os outros são pouco prováveis. Isso é condizente com uma região ser muito povoada pela mesma classe.
- O índice de Gini confome vimos em sala de aula, pode ser interpretado como proveniente de uma expansão em série de Taylor do logaritmo na entropia, sendo portanto, uma simplificação dessa quantidade, com um custo computacional menor.