Mineração de Dados

Aula 9: Bagging, Random Forests e Boosting

Rafael Izbicki

Combinando Predições

Imagine que, em um contexto de regressão, temos dois preditores para Y, $g_1(\mathbf{x})$ e $g_2(\mathbf{x})$.

O risco destes (condicional em $\mathbf{x}$) é dado por

$$\mathbb{E}\left[(Y - g_1(\mathbf{x}))^2|\mathbf{x}\right] \quad \text{e} \quad \mathbb{E}\left[(Y - g_2(\mathbf{x}))^2|\mathbf{x}\right].$$

Considere o estimador $g(\mathbf{x}) = (g_1(\mathbf{x}) + g_2(\mathbf{x}))/2.$. Temos

$$\begin{aligned}
&\mathbb{E}\left[(Y - g(\mathbf{x}))^2 |\mathbf{x}\right] = \\
&= \mathbb{V}[Y|\mathbf{x}] + \frac{1}{4}\left(\mathbb{V}[g_1(\mathbf{x}) + g_2(\mathbf{x})|\mathbf{x}]\right) + \\
&+ \left(\mathbb{E}[Y|\mathbf{x}] - \frac{\mathbb{E}[g_1(\mathbf{x})|\mathbf{x}] + \mathbb{E}[g_2(\mathbf{x})|\mathbf{x}]}{2}\right)^2
\end{aligned}$$

Assim, se g_1 e g_2 são não correlacionados, não viesados e têm mesma variância,

$$\mathbb{E}\left[(Y - g(\mathbf{x}))^2 \,|\mathbf{x}\right] = \mathbb{V}[Y|\mathbf{x}] + \frac{1}{2}\mathbb{V}[g_i(\mathbf{x})|\mathbf{x}] \leq \mathbb{E}\left[(Y - g_i(\mathbf{x}))^2|\mathbf{x}\right],$$

$i = 1, 2.$

Assim, é melhor se utilizar o estimador combinado g do que usar g_1 ou g_2 separadamente.

Random Forests/Bagging: usar essas ideias para melhorar predições em árvores de predição (que, em geral, não possuem bom poder preditivo)

Criaremos B árvores, e combinaremos seus resultados para melhorar o poder preditivo de cada árvore individual.

Para criar árvores próximas de não-viesadas, não iremos podá-las.

Bagging

Ideia: Criamos B amostras bootstrap da amostra original (i.e., amostra com reposição da amostra original).

Para cada um delas, criamos uma árvore (não a podamos).

Seja $g^b(\mathbf{x})$ a função de predição obtida segundo a b-ésima árvore.

A função de predição dada pelo bagging é dada por

$$g(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} g^b(\mathbf{x})$$

no caso de regressão, e, para classificação,

$$g(\mathbf{x}) = moda_b g^b(\mathbf{x})$$

A ideia é que combinando várias funções de predição obtemos um estimador com variância menor.

As árvores não são podadas para diminuir o viés de cada função de predição.

Perdemos bastante da interpretação de árvores

Uma medida de importância para cada covariável: a média de quanto ela foi importante em cada árvore.

Random Forests - Florestas Aleatórias

Mesma ideia de bagging, mas cada nó só pode escolher uma dentre $m < p$ covariáveis.

O subconjunto de covariáveis é escolhido aleatoriamente para cada nó.

A ideia é diminuir a correlação entre os diferentes g^b's (se todas as variáveis podem ser usadas como no bagging, os g^b's tendem a ser muito próximos uns dos outros, de modo que tem covariância alta e, portanto, a variância do estimador combinado não é tão menor.).

Logo, aumenta-se o viés em troca de uma diminuição da variância.

m pode ser escolhido por validação cruzada. Em geral $m \approx \sqrt{p}$ possui boa performance.

Aquivo R

Boosting para regressão

Ideia: construimos $g(\mathbf{x})$ incrementalmente. Começamos com $g(\mathbf{x}) \equiv 0$, que possui alto viés e baixa variância. Aos poucos, diminuimos o viés e aumentamos a variância. Algoritmo:

1. Definimos $g(\mathbf{x}) \equiv 0$ e $r_i = y_i \; \forall i$.

2. Para $b = 1, \ldots, B$:

   a. Ajustamos uma árvore com d folhas para $(\mathbf{x}_1, r_1), \ldots, (\mathbf{x}_n, r_n)$. Seja $g^b(\mathbf{x})$ sua respectiva função de predição.

   b. Atualizamos g e os resíduos: $g(\mathbf{x}) \leftarrow g(\mathbf{x}) + \lambda g^b(\mathbf{x})$ e $r_i \leftarrow Y_i - g(\mathbf{x})$.

3. Saída: modelo final $g(\mathbf{x})$

Tuning parameters: B, d e λ. Tipicamente λ é pequeno (e.g., 0.001), $B \approx 1000$ e d é da ordem de 2 ou 4. Multiplicação por λ é usada para que a variância não seja grande.

Boosting para classificação

Assumindo que $y_i \in \{-1, 1\}$:

1. Inicialize os pesos $w_1 = \ldots = w_n = \frac{1}{n}$
2. Para $b = 1, \ldots, B$:
   2.1 Ajuste um classificador $g_b(\mathbf{x})$ para a amostra de treinamento usando os pesos $w_1, \ldots, w_n$
   2.2 Calcule o erro $er_b = \frac{\sum_{i=1}^n w_i \mathbb{I}(y_i \neq g_b(\mathbf{x}_i))}{\sum_{i=1}^n w_i}$
   2.3 Calcule $\alpha_b = \log((1 - er_b)/er_b)$
   2.4 Atualize $w_i \leftarrow w_i \exp\left(\alpha_b \mathbb{I}(y_i \neq g_b(\mathbf{x}_i))\right)$, $i = 1, \ldots, n$
3. Retornamos o modelo final $g(\mathbf{x}) = \text{sinal}\left(\sum_{b=1}^{B} \alpha_b g_b(\mathbf{x})\right)$

Código R (spamForests.r)