Gradient Boosting Practice: Poisson Response

David S. Rosenberg

Suppose we’re trying to predict a distribution of count from some input co-
variates. The simplest distribution in this situation is the Poisson distribution:

\[ p(k; \lambda) = \frac{e^{-\lambda} \lambda^k}{k!} \]

on \( k = 0, 1, 2, 3, \ldots \lambda \in (0, \infty) \).

1 Linear Conditional Probability Model

- **Input:** \( x \in \mathbb{R}^d \).
- **Output:** \( y \in \{0, 1, 2, \ldots\} \)
- **Data:**
  \[
  \mathcal{D} = ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathbb{R}^d \times \{0, 1, 2, \ldots\})^n
  \]
  assume is sampled i.i.d. from some distribution \( P_{X \times Y} \).
- **Action:** \( \lambda \in (0, \infty) \), where \( \lambda \) is the parameter of a Poisson distribution.

We’ve got to map input \( x \) to action \( \lambda \) in our action space, which is \( (0, \infty) \).

\[
  x \mapsto w^T x \quad \mapsto \lambda \quad \in (0, \infty)
\]

To map the score \( s \) into our action space, we could use the transfer function \( \psi(s) = \exp(s) \). Then

\[
  \lambda = \exp(w^T x).
\]
So if we predict $\lambda$, what’s the probability of an observed count $k$ for input vector $x$?

$$p(y = k \mid x; w) = \frac{e^{-\lambda(x)} \lambda(x)^k}{k!} = e^{-\exp(w^T x)} \frac{[\exp(w^T x)]^k}{k!}.$$ 

The conditional likelihood for particular example $(x_i, y_i)$, (where $y_i$ is a count) is

$$p(y = y_i \mid x_i; w) = e^{-\exp(w^T x_i)} \frac{[\exp(w^T x_i)]^{y_i}}{y_i!}.$$ 

Easier to work with the log:

$$\log p(y = y_i \mid x_i; w) = -\exp(w^T x_i) + y_i w^T x_i - \log(y_i!$$

What do we need to find to fit this model? $w$. Our strategy is to use maximum log-likelihood:

$$\log L_D(w) = \log p(D; w)$$

$$= \sum_{i=1}^{n} \log p(y = y_i \mid x_i; w)$$

$$= \sum_{i=1}^{n} [-\exp(w^T x_i) + y_i w^T x_i - \log(y_i!)]$$

So find $w$ maximizing this log-likelihood and we’re done. Can use standard gradient based methods.

\section*{2 Nonlinear approach}

In a nonlinear approach, we’ll replace the linear score function $s = w^T x$ with a nonlinear function $s = f(x)$:

$$x \mapsto f(x) \mapsto \lambda \in (0, \infty).$$
Again, we can use the transfer function $\psi(s) = \exp(s)$. So

$$\lambda = \exp\left(f(x)\right).$$

For score function $f$, the probability of $y_i \mid x_i$ is:

$$p(y = y_i \mid x_i; f) = \frac{e^{-\exp(f(x_i))} \left[\exp\left(f(x_i)\right)\right]^{y_i}}{y_i!}.$$

Easier to work with the log:

$$\log p(y = y_i \mid x_i; f) = -\exp\left(f(x_i)\right) + y_i f(x_i) - \log(y_i!)$$

Somehow we want to find a function $f$ that gives high log-likelihood to our observed data:

$$\log L_D(w) = \sum_{i=1}^{n} \left[ -\exp\left(f(x_i)\right) + y_i f(x_i) - \log(y_i!) \right]$$

### 3 Gradient Boosting Approach

Let’s differentiate $\log p(y = y_i \mid x_i; f)$ w.r.t. $f(x_i)$:

$$\frac{\partial}{\partial f(x_i)} \log p(y = y_i \mid x_i; f) = -\exp\left(f(x_i)\right) + y_i$$

Now differentiating the full log-likelihood is

$$\frac{\partial}{\partial f(x_i)} \left[ \log L_D(f) \right] = \frac{\partial}{\partial f(x_i)} \left[ -\exp\left(f(x_i)\right) + y_i f(x_i) - \log(y_i!) \right]$$

$$= -\exp\left(f(x_i)\right) + y_i$$

So optimal unconstrained step direction for changing the vector of evaluations $f = (f(x_1), \ldots, f(x_n))$ is

$$-g = (-y_1 + \exp(f(x_1)), \ldots, -y_n + \exp(f(x_n)))$$

Fix some base hypothesis space $\mathcal{H}$ of functions $h : \mathbb{R}^d \rightarrow \mathbb{R}$. Then, our actual step direction will be the $h \in \mathcal{H}$ that best fits $-g$ in the least squares sense:

$$\arg \min_{h \in \mathcal{H}} \sum_{i=1}^{n} (-g_i - h(x_i))^2$$

$$= \arg \min_{h \in \mathcal{H}} \sum_{i=1}^{n} \left(\left[-y_i + \exp\left(f(x_i)\right)\right] - h(x_i)\right)^2$$

So to recap:
1. Up to this point, our score function is $f$.

2. We want to improve $f$.

3. The optimal step direction for $f(x_i)$ is $-y_i + \exp(f(x_i))$. We can evaluate this. It’s a real number.

4. So we have a bunch of $(x_i, -g_i)$ pairs that we will use regression over $\mathcal{H}$ to fit.

Then we add something like $0.1h$ to $f$ and repeat.