CS195-5, Lecture 3: Derivations and notes
by Greg Shakhnarovich `gregory@cs`

# Slides 4-5

By definition of expectation, the expected loss with respect to the joint distribution of $\mathbf{x}_0$ and $y_0$ is

$$E_{(\mathbf{x}_0,y_0)\sim p(\mathbf{x},y)}\left[(f(\mathbf{x}_0)-y_0)^2\right] \;=\; \int_{\mathbf{x}_0}\int_{y_0}(f(\mathbf{x}_0)-y_0)^2\,p(\mathbf{x}_0,y)dy_0 d\mathbf{x}_0. \quad (1)$$

By definition of conditional probability,

$$p(a|b)=\frac{p(a,b)}{p(b)},$$

so that $p(\mathbf{x}_0,y_0)=p(y_0|\mathbf{x}_0)p(\mathbf{x}_0)$. Plug that into (1):

$$E_{(\mathbf{x}_0,y_0)\sim p(\mathbf{x},y)}\left[(f(\mathbf{x}_0)-y_0)^2\right] \;=\; \int_{\mathbf{x}_0}\int_{y_0}(f(\mathbf{x}_0)-y_0)^2\,p(y|\mathbf{x}_0)p(\mathbf{x}_0)dy_0 d\mathbf{x}_0 \quad (2)$$

$$= \int_{\mathbf{x}_0}\left\{\int_{y_0}(f(\mathbf{x}_0)-y_0)^2\,p(y|\mathbf{x}_0)dy_0\right\}p(\mathbf{x}_0)d\mathbf{x}_0. \quad (3)$$

We could switch the order of $dy_0$ and $p(\mathbf{x}_0)$ since $p(\mathbf{x}_0)$ does not depend on $y_0$.

Let us denote the expression within the curly brackets in (3)

$$g(\mathbf{x}_0) \;\triangleq\; \int_{y_0}(f(\mathbf{x}_0)-y_0)^2\,p(y|\mathbf{x}_0)dy_0.$$

Note that this is a function of $x_0$, but *not* of $y_0$ ($y_0$ is "integrated out"). This function $g$ is the *conditional expectation of $y_0$ given* $\mathbf{x}_0$:

$$g(\mathbf{x}_0)=E_{p(y_0|\mathbf{x}_0)}\left[(f(\mathbf{x}_0)-y_0)^2\mid\mathbf{x}_0\right]$$

Intuitively, it tells us, for a given value of $\mathbf{x}_0$, what is the "average" value of $y_0$ we should expect for that $\mathbf{x}_0$. Plugging this definition of $E_{p(y_0|\mathbf{x}_0}$ into (3) produces the expression on the top of slide 5.

Now, the quantity in (3) is

$$\int_{\mathbf{x}_0} g(\mathbf{x}_0)p(\mathbf{x}_0)d\mathbf{x}_0 \;=\; E_{p(\mathbf{x}_0)}\left[g(\mathbf{x}_0)\right], \tag{4}$$

the equality is due to definition of expectation. This averages the value of $g(\mathbf{x}_0)$ for all possible values of $\mathbf{x}_0$, weighted by their probability. Putting this all together we get the form on the bottom of slide 4:

$$E_{(\mathbf{x}_0,y_0)\sim p(\mathbf{x},y)}\left[(f(\mathbf{x}_0)-y_0)^2\right] \;=\; E_{\mathbf{x}_0\sim p(\mathbf{x})}\left[E_{y_0\sim p(y|\mathbf{x})}\left[(f(\mathbf{x}_0)-y_0)^2 \,|\, \mathbf{x}_0\right]\right] \tag{5}$$

# Slide 11

We are looking at

$$E_{p(y|\mathbf{x})}\left[f(\mathbf{x};\mathbf{w}) \;+\; \nu \,|\, \mathbf{x}\right] \tag{6}$$

By *linearity* of expectation (easy to derive directly from the definition of expectation as an integral), for any quantities $A$ and $B$

$$E\left[A+B\right] = E\left[A\right] + E\left[B\right],$$

hence

$$E_{p(y|\mathbf{x})}\left[f(\mathbf{x};\mathbf{w}) \;+\; \nu \,|\, \mathbf{x}\right] \;=\; E_{p(y|\mathbf{x})}\left[f(\mathbf{x};\mathbf{w})\right] \;+, E_{p(y|\mathbf{x})}\left[\nu\right].$$

Under our model, $f(\mathbf{x};\mathbf{w})$ is deterministic (non-random). Simply put, for a given $\mathbf{x}$, it's just a number. Therefore,

$$E_{p(y|\mathbf{x})}\left[f(\mathbf{x};\mathbf{w})\right] \;=\; f(\mathbf{x};\mathbf{w}).$$

The noise $\nu$ is random, with distribution $p(\nu)$. It does not, under our assumptions, depend on $\mathbf{x}$; it doesn't depend on anything. Therefore, it's expectation, whether conditioned on $\mathbf{x}$ or not, is simply given by $E_{p(\nu)}\left[\nu\right]$ - and we get the result on the slide.