IntroML - Lecture Notes Week 2.5

Ruben Schenk, ruben.schenk@inf.ethz.ch

March 22, 2022

1 Math Recap

1.1 Derivatives

$f(x) = Ax$: Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be defined as $f(x) = Ax$ for some matrix $A \in \mathbb{R}^{m \times m}$. As f itself is linear, the derivative at any point such as x_0 coincides with f itself:

$$Df(x_0) = A$$

$f(x) = w^T x$: Let $f : \mathbb{R}^n \to \mathbb{R}$ be defined as $f(x) = w^T x$. Similar to the previous case, f is linear, hence its derivative is itself:

$$Df(x_0) = w^T$$

Note: The vector w is called the **gradient** of f at x_0.

$g(x) = x^T Ax$: Let $g : \mathbb{R}^n \to \mathbb{R}$ be defined as $g(x) = x^T Ax$. The goal is to compute the gradient of g at x. Observe that

$$g(x) = \sum_{i,j=1}^{n} a_{ij} x_i x_j.$$

The derivative of g is going to be a $1 \times n$ matrix with elements beingequal to the partial derivatives of g:

$$Dg(x) = [\frac{\partial g}{\partial x_1} \cdots \frac{\partial g}{\partial x_n}]$$

By computing we get:

$$\frac{\partial g}{\partial x_i} = \sum_{j=1}^{n} a_{ij} x_j + \sum_{j=1}^{n} a_{ji} x_j$$

Packing the partial derivatives gives:

$$Dg(x) = ((A + A^T)x)^T, \quad \nabla g(x) = (A + A^T)x$$

Specifically, if A is *symmetric*, we have $\nabla g(x) = 2Ax$.

1.2 Probability

Lemma: Let X, Y be random variables, and f, $g : \mathbb{R} \to \mathbb{R}$ two functions. Then:

$$\mathbb{E}[f(X)g(y) \mid X] = f(X)\mathbb{E}[g(Y) \mid X].$$

Lemma: Let X, Y be two random variables. Then:

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y \mid X]].$$

The two lemmas can be used to show that:

$$\mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(X \mid Y)] + \mathrm{Var}(\mathbb{E}[Y \mid X]).$$