Painless Unsupervised Learning with Features

Taylor Berg-Kirkpatrick Alexandre Bouchard-Côté John DeNero Dan Klein
Computer Science Division
University of California at Berkeley
Berkeley, CA 94720

1 Proof of the gradient

We first prove the following lemma:

Lemma 1 *If ϕ, ψ are real-valued functions such that:*

1. $\phi(\boldsymbol{x}_0) = \psi(\boldsymbol{x}_0)$ *for some* $\boldsymbol{x}_0$,
2. $\phi(\boldsymbol{x}) \leq \psi(\boldsymbol{x})$ *on an open set* S *containing* $\boldsymbol{x}_0$,
3. ϕ *and* ψ *are differentiable at* $\boldsymbol{x}_0$,

then $\nabla\psi(\boldsymbol{x}_0) = \nabla\phi(\boldsymbol{x}_0)$.

Proof: Without loss of generality, ϕ, ψ are univariate functions with $\phi(x_0) = \psi(x_0) = 0$, and $x_0 = 0$.

Let $\delta = \psi'(x_0) - \phi'(x_0)$ and consider a sequence $a_n > 0$ converging to zero with $a_n \in S$. We have:

$$\lim_{n\to\infty} \frac{\psi(a_n) - \phi(a_n)}{a_n} = \delta,$$

and since the numerator and denominator are both positive for all n, we conclude that $\delta \geq 0$.

By doing the same argument with a sequence $b_n < 0$ converging to zero, we get that $\delta \leq 0$, hence the derivatives are equal. ∎

Theorem 2 *Algorithm 2 computes the gradient of the log marginal likelihood:*

$$\nabla L(\mathbf{w}) = \nabla \ell(\mathbf{w}, \mathbf{e})$$

.

Proof: To prove the theorem, we introduce the following notation:

$$H(\mathbf{w}) = -\sum_{\mathbf{z}} P_{\mathbf{w}}(\mathbf{Z} = \mathbf{z} | \mathbf{Y} = \mathbf{y}) \log P_{\mathbf{w}}(\mathbf{Z} = \mathbf{z} | \mathbf{Y} = \mathbf{y}),$$

and we set:

$$\begin{aligned} \psi(\mathbf{w}) &= L(\mathbf{w}) \\ \phi(\mathbf{w}) &= \ell(\mathbf{w}, \mathbf{e}) + H(\mathbf{w}_0). \end{aligned}$$

If we can show that ψ, ϕ satisfy the conditions of the lemma, we are done since the second term of ϕ depends on $\mathbf{w}_0$, but not on $\mathbf{w}$.

Property (3) can be easily checked, and property (3) follows from Jensen's inequality. To show property (1), note that:

$$\begin{aligned} \phi(\mathbf{w}_0) &= \sum_{\mathbf{z}} P_{\mathbf{w}_0}(\mathbf{Z} = \mathbf{z} | \mathbf{Y} = \mathbf{y}) \log \frac{P_{\mathbf{w}_0}(\mathbf{Z} = \mathbf{z}, \mathbf{Y} = \mathbf{y})}{P_{\mathbf{w}_0}(\mathbf{Z} = \mathbf{z} | \mathbf{Y} = \mathbf{y})} - \kappa ||\mathbf{w}_0||_2^2 \\ &= \sum_{\mathbf{z}} P_{\mathbf{w}_0}(\mathbf{Z} = \mathbf{z} | \mathbf{Y} = \mathbf{y}) \log P_{\mathbf{w}_0}(\mathbf{Y} = \mathbf{y}) - \kappa ||\mathbf{w}_0||_2^2 \\ &= L(\mathbf{w}_0). \end{aligned}$$

■