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 $\phi, \psi$ are real-valued functions such that:

1. $\phi(x_0) = \psi(x_0)$ for some $x_0$.
2. $\phi(x) \leq \psi(x)$ on an open set $S$ containing $x_0$,
3. $\phi$ and $\psi$ are differentiable at $x_0$,

then $\nabla \psi(x_0) = \nabla \phi(x_0)$.

Proof: Without loss of generality, $\phi, \psi$ 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.  

\[ \blacksquare \]

1
Theorem 2 Algorithm 2 computes the gradient of the log marginal likelihood:

\[ \nabla L(w) = \nabla \ell(w, e) \]

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

\[ H(w) = -\sum_z P_w(Z = z|Y = y) \log P_w(Z = z|Y = y), \]

and we set:

\[ \psi(w) = L(w) \]
\[ \phi(w) = \ell(w, e) + H(w_0). \]

If we can show that \( \psi, \phi \) satisfy the conditions of the lemma, we are done since the second term of \( \phi \) depends on \( w_0 \), but not on \( w \).

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

\[ \phi(w_0) = \sum_z P_{w_0}(Z = z|Y = y) \log \frac{P_{w_0}(Z = z, Y = y)}{P_{w_0}(Z = z|Y = y)} - \kappa||w_0||^2 \]
\[ = \sum_z P_{w_0}(Z = z|Y = y) \log P_{w_0}(Y = y) - \kappa||w_0||^2 \]
\[ = L(w_0). \]