Derivation of variance-weighted Bernoulli equations

Here we derive the likelihood functions for variance-weighted Bernoulli and mixture-of-Bernoulli models, Eqns. 4 and 5 in Section 3.2 of the submitted paper,

Recall that each foreground shape mask is modeled as a set $\mathbf{x} = (x_1, \ldots, x_D)^T$ of D binary variables. If each of the pixels x_d of $\mathbf{x}$ is modeled as a Bernoulli variable with mean μ_d, then assuming conditional independence among pixels, the joint likelihood function is given by

$$p(\mathbf{x}|\mu) = \prod_{d=1}^{D} \mu_d^{x_d} (1-\mu_d)^{(1-x_d)} \qquad (1)$$

We now associate with each pixel x_d a nonnegative weight v_d computed as the variance of that pixel across all the training patterns. We can think of these weights as representing the importance of each pixel, with higher weights meaning more important, in order to make the model spend more effort explaining the higher-variance parts of the shape.

To see how to incorporate these weights into the likelihood function, imagine the weights were integers. We could then treat them as replication factors, saying how many times to duplicate each pixel to increase its influence. Consider a simple case where $D = 2$; Eqn. 1 hence represents the joint likelihood for two pixels x_1 and x_2 so that

$$p(\mathbf{x}|\mu) = p(x_1, x_2|\mu) = p(x_1|\mu_1) p(x_2|\mu_2) \ .$$

Now, if we thought that it was twice as important to explain pixel x_2 as it was to explain x_1, we could duplicate x_2 to get two copies of it, as opposed to only one copy of x_1, so the joint likelihood would become

$$\begin{aligned} p(x_1, x_2, x_2|\mu) &= p(x_1|\mu_1) p(x_2|\mu_2) p(x_2|\mu_2) \\ &= [p(x_1|\mu_1)]^1 [p(x_2|\mu_2)]^2 \end{aligned}$$

We could achieve the same effect by setting weight $v_1 = 1$ and $v_2 = 2$, and writing

$$p(x_1, x_2|\mu, v_1, v_2) = [p(x_1|\mu_1)]^{v_1} [p(x_2|\mu_2)]^{v_2} \ .$$

In general, allowing the weights to be noninteger values (as long as they are nonnegative), the joint likelihood function of a set of weighted Bernoulli variables can be written as

$$p(\mathbf{x}|\mu) = \prod_{d_1}^{D} \mu_d^{x_d v_d} (1-\mu_d)^{(1-x_d) v_d} \qquad (2)$$

which is what we proposed in Eqn. 4 in Section 3.2 of the submitted paper.

Weighted Bernoulli mixture model

For a collection $\mathbf{X} = \{\mathbf{x}_i, \ldots, \mathbf{x}_N\}$ of N training shape patterns, taking the weighted Bernoulli variables above as the components of a mixture model leads to our weighted Bernoulli mixture model. The joint likelihood of the model is thus

$$p(\mathbf{X}|\mu, \pi) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \left\{ \prod_{d=1}^{D} {\mu_{kd}}^{v_d x_{nd}} (1-\mu_{kd})^{v_d (1-x_{nd})} \right\} \qquad (3)$$

where $\pi = \{\pi_1, \ldots, \pi_K\}$ are the component mixing weights.

The rest of the derivation for the log-likelihood function (Eqn. 5 in the submitted paper) and parameter estimation proceeds in the standard way by hypothesizing a set of latent variables representing the component membership of each training sample, relaxing that to be a soft assignment computed as expected values, and iterating expectation and maximization steps within an EM algorithm. A similar derivation (without weights) can be found in Bishop.

References C. Bishop, *Pattern Recognition and Machine Learning,* Chapter 9, Mixture Models and EM, Springer, 2006.