Slides 6

Problem 2.1

The bernouli likelihood function, defined for random variable $t \in -1, 1$, takes the form:

$$p(t = 1) = p$$
$$p(t = -1) = 1 - p$$

If we now consider a sigmoid function $\sigma(u)$ which gives us a bernouli probability, $p \in [0, 1]$, we then get the following:

$$p(t = 1) = \sigma(u)$$
$$p(t = -1) = 1 - \sigma(u)$$

If we consider the situation where $\sigma(u)$ is the logistic function:

$$p(t = 1) = \frac{1}{1 + e^{-u}}$$
$$p(t = -1) = 1 - \frac{1}{1 + e^{-u}}$$

The sigmoid/logistic function has the special property that in addition to have an output bounded between $[0, 1]$, it's first derivative is also symmetric around 0. This is to say that from $[0, \infty]$, the function moves towards 1 at the exact same rate it moves towards 0 over the range $[0, -\infty]$. This implies the following:

$$\sigma(u) = 1 - \sigma(-u)$$

Using this, we can rewrite our previous pair of equations:

$$p(t = 1) = \frac{1}{1 + e^{-u}}$$
$$p(t = -1) = \frac{1}{1 + e^{u}}$$

Or better still:

$$p(t) = \frac{1}{1 + e^{-tu}}$$

If we then plug the linear model $\boldsymbol{w}^T\phi(x)+b$ in for u, we find exactly the large margin classifier with the logistic loss function:

$$p(t) = \frac{1}{1+\exp\Big\{-t(\boldsymbol{w}^T\phi(x)+b)\Big\}}$$

Problem 2.2

The MAP estimator is given, as per usual:

$$w_{MAP} = \underset{w}{\text{argmax}}\, p(\boldsymbol{t}|\boldsymbol{w}, X)p(\boldsymbol{w})$$

Our $\boldsymbol{t}$ observed variables are bernouli $t \in -1, 1$, and the linear model we are applying on our data to map it to our bernouli predictions is the logistic function. Our likelihood function $p(\boldsymbol{t}|\boldsymbol{w}, X)$, is therefore exactly what we found in 2.1, while our prior on $\boldsymbol{w}$ is Gaussian. We include only the terms that are dependent on $\boldsymbol{w}$ and assume independence between rows of our design matrix:

$$w_{MAP} = \underset{w}{\text{argmax}} \prod_n \left(\frac{1}{1+\exp\Big\{-t_n(\boldsymbol{w}^T\phi(x_n)+b)\Big\}}\right) * \exp\Big\{-\frac{1}{2}\boldsymbol{w}^T\lambda N\,\boldsymbol{w}\Big\}$$

$$w_{MAP} = \underset{w}{\text{argmax}} \sum_n \Big(-\log\Big(1+\exp\Big\{-t_n(\boldsymbol{w}^T\phi(x_n)+b)\Big\}\Big)\Big) - \frac{\lambda N}{2}\boldsymbol{w}^T\boldsymbol{w}$$

$$w_{MAP} = \underset{w}{\text{argmin}} \sum_n \log\Big(1+\exp\Big\{-t_n(\boldsymbol{w}^T\phi(x_n)+b)\Big\}\Big) + \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w}$$

Slides 7

Problem 2

The probability that a new point belongs to any given mixture k is the joint probability of that data and the latent variable that points to that particular mixture. We decompose this into the conditional probability of the data

given that particular mixture multiplied by the prior on the probability of that mixture. To coerce this into a probability, we normalize by the sum of the probabilities that the point belongs to each mixture k:

$$\Pi_k = \frac{p(x_{n+1}, z_k)}{\sum_i^K p(x_{n+1}|z_i)}$$
$$\Pi_k = \frac{p(x_{n+1}|z_k)p(z_k)}{\sum_i^K p(x_{n+1}|z_i)}$$

We will focus on determining the formula for the numerator, for one particular k, as this is simply repeated for all k's to create a vector of probabilities, and the denominator follows directly from the numerator. The latent variable z_k corresponds to the mean μ_k and precision Q_k for that particular guassian, so we compute the probability that x_{n+1} came from that gaussian, multiplied by whatever our prior probability was on that gaussian:

$$p(x_{n+1}|z_k)p(z_k)$$
$$\frac{exp\Big\{ -\frac{1}{2}(x_{n+1} - \mu_k)^T Q(x_{n+1} - \mu_k)\Big\}}{\sqrt{|2\pi Q^{-1}|}} * p(z_k)$$

Problem 5

Robust Regression

Assuming ν is fixed and applying bayes theorum:

$$p(\eta_n|t_n, \nu) = \frac{p(t_n|\nu, \eta_n)p(\eta_n|\nu)}{p(t_n)}$$
$$p(\eta_n|t_n, \nu) = \frac{p(t_n, \eta_n|\nu)}{p(t_n)}$$

In robust regression we model the observed data, $\boldsymbol{t}$, as gaussian, but where the gaussian for every observation has a variance derived from a latent variable η_n which itself is a random variable picked from a gamma distribution. As shown above, this can be rewritten as the joint distribution of two random variables, t_n and ν_n, which coincides with the definition of the Normal-Gamma distribution.

Logistic and Probit Models

z_n is given as a univariate random variable that consists of a linear combination of $\phi(\boldsymbol{x}_n)^T\boldsymbol{w}$ plus noise, whose distribution is either gaussian or logistic.

$$p(z_n|\boldsymbol{w}, \boldsymbol{x}_n) \sim \mathcal{N}(\phi(\boldsymbol{x}_n)^T\boldsymbol{w}, \sigma^2)$$

$$p(z_n|\boldsymbol{w}, \boldsymbol{x}_n) \sim Logistic(\phi(\boldsymbol{x}_n)^T\boldsymbol{w}, \frac{\sqrt{3}\sigma}{\pi})$$

t_n is given to us as $t_n = 1[z_n > 0]$, meaning in terms of z_n:

$$p(z_n \leq 0|t_n = 1) = 0$$
$$p(z_n > 0|t_n = 1) = 1$$
$$p(z_n \leq 0|t_n = 0) = 1$$
$$p(z_n > 0|t_n = 0) = 0$$

Which would imply that conditioning $z_n|\boldsymbol{w}, \boldsymbol{x}_n$ on t_n as well, provides us no additional information beyond the sign of z_n. The mean of our distributions is deterministically decided by the linear combination of data and weights, $\phi(\boldsymbol{x}_n)^T\boldsymbol{w}$, however we now have a truncated for for this distribution, as we know the sign of the outcome variable. This implies:

$$p(z_n|\boldsymbol{w}, \boldsymbol{x}_n, t_n = 1) = p(z_n|\boldsymbol{w}, \boldsymbol{x}_n, z_n > 0)$$
$$p(z_n|\boldsymbol{w}, \boldsymbol{x}_n, t_n = 0) = p(z_n|\boldsymbol{w}, \boldsymbol{x}_n, z_n \leq 0)$$

We understand a truncated distribution to be the distribution that would follow if we threw away all values that were outside of the truncation line, and re-drew from the distribution. Following from that definition of a truncated distribution, it is known the distribution can be written as:

$$p(y|y > 0) = \frac{p(y)}{1 - p(y \leq 0)}$$
$$p(y|y \leq 0) = \frac{p(y)}{p(y \leq 0)}$$

The intuition behind this is that we want to scale the bounded distribution UP by the amount of the unbounded distribution found outside of the boundary, which equates to dividing by the percent of the unbounded distribution found within the boundary. We will define a function, $g(x)$ to be the CDF any random variable x evaluated at 0:

$$g(x) = \Phi_x(0) = p(x \leq 0)$$

This becomes convenient to rewrite to include both possible values of t and here we have our density of $z_n|\boldsymbol{w}, \boldsymbol{x}_n, t_n$:

$$p(z_n|\boldsymbol{w}, \boldsymbol{x}_n, t_n) = \frac{p(z_n|\boldsymbol{w}, \boldsymbol{x}_n)}{g(z_n|\boldsymbol{w}, \boldsymbol{x}_n)(g(z_n|\boldsymbol{w}, \boldsymbol{x}_n)^{-1} - g(z_n|\boldsymbol{w}, \boldsymbol{x}_n))^{t_n}} \tag{1}$$

Now we define $g(z_n|\boldsymbol{w}, \boldsymbol{x}_n)$ and $p(z_n|\boldsymbol{w}, \boldsymbol{x}_n)$ for both guassian and logistic densities, which will allow us to use (1) to solve for $z_n|\boldsymbol{w}, \boldsymbol{x}_n, t_n$ for probit and logit models, respectively:

ALTERNATE

We will consider a general case of K discrete possible values for t, which will easily hold true for our binomial case, and will be shown to be a more general expression of the truncated distribution in the currently-considered binomial case. We can marginalize over t:

$$p(z_n|\boldsymbol{w}, \boldsymbol{x}_n) = \sum_k p(z_n|\boldsymbol{w}, \boldsymbol{x}_n, t_k)p(t_k|\boldsymbol{w}, \boldsymbol{x}_n)$$

Expanding the sum over possible values of t, and gathering terms, we come to:

$$p(z_n|\boldsymbol{w}, \boldsymbol{x}_n, t_n) = \frac{p(z_n|\boldsymbol{w}, \boldsymbol{x}_n) - \sum_{k\neq n}^{K} p(z_n|\boldsymbol{w}, \boldsymbol{x}_n, t_k)p(t_k|\boldsymbol{w}, \boldsymbol{x}_n)}{p(t_n|\boldsymbol{w}, \boldsymbol{x}_n)}$$

We gather the marginal distributions into their joint distribution as follows:

$$p(z_n|\boldsymbol{w}, \boldsymbol{x}_n, t_n) = \frac{p(z_n|\boldsymbol{w}, \boldsymbol{x}_n) - \sum_{k\neq n}^{K} p(z_n, t_k|\boldsymbol{w}, \boldsymbol{x}_n)}{p(t_n|\boldsymbol{w}, \boldsymbol{x}_n)} \tag{2}$$

We can begin to see some things in the above equation that match our intuition, and similarly match the definition of a truncated distribution, where we have turned a continous distribution function into a piece-wise function that returns 0 over certain intervals. For any z_n that deterministically gives us t_k:

$$p(z_n|\boldsymbol{w}, \boldsymbol{x}_n) = p(z_n, t_k|\boldsymbol{w}, \boldsymbol{x}_n)$$

Together with our (2), this gives us the desired result that for any z_n that deterministically gives us t_k, where $t_k \neq t_n$:

$$p(z_n|\boldsymbol{w}, \boldsymbol{x}_n, t_n) = 0$$

This gives us a probability function that will return 0 for any z_n outside of the range of the t_n it is conditioned upon, and for every other value of z_n will return:

$$p(z_n|\boldsymbol{w}, \boldsymbol{x}_n, t_n) = \frac{p(z_n|\boldsymbol{w}, \boldsymbol{x}_n)}{p(t_n|\boldsymbol{w}, \boldsymbol{x}_n)} \tag{3}$$

We will show that $p(t_n|\boldsymbol{w}, \boldsymbol{x}_n)$ is equal to the probability given by the CDF of $p(z_n|\boldsymbol{w}, \boldsymbol{x}_n)$ of z_n being within the range of t_n, which relates back to the truncated distribution over z_n, which attempts to scale up the probabilities within the range by the amount of total distributed mass outside the range. We now plug in the gaussian and logistic distribution functions for the probit and logit models, respectively:

Gaussian Mixture Models

See probelm 2?

Slides 8