CS281 Section #2 Notes, v 1.0

Due: never

1 Gaussian Distribution

1.1 Review

MVN: Let X be a D-dimensional MVN random vector with mean μ and covariance matrix $\mathbf{\Sigma}$, denoted $X \sim \mathcal{N}(\mu, \mathbf{\Sigma})$. Then the pdf of X is

$$p(x) = (2\pi)^{-D/2} |\mathbf{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}(x-\mu)^T \mathbf{\Sigma}^{-1}(x-\mu)\right]$$

MLE of Parameters: We can find the MLE for the parameters of an MVN like this:

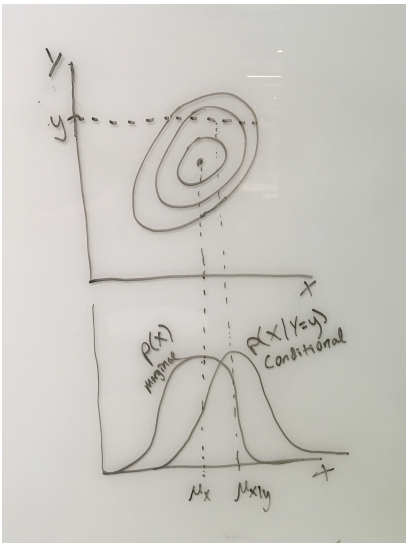
$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i = \bar{\mathbf{x}} \text{(sample mean)}$$

$$\hat{\mathbf{\Sigma}} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T = \frac{1}{N}(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^T) - \bar{\mathbf{x}}\bar{\mathbf{x}}^T$$

So far, this should all look a lot like what you've seen with univariate Gaussians.

Marginal Probability: In the 2-D Gaussian below, we can compute the marginal distribution of X (written $p(X)$) or Y (written $p(Y)$). This picture shows the marginal distribution of X, which corresponds to slicing the distribution at the mean value of Y.

Conditional Probability: We can also compute the conditional probability of X given that Y takes a specific value y (written $p(X|Y=y)$), or the conditional probability of Y given that X takes a specific value x (written $p(Y|X=x)$). The plot below shows $p(X|Y=y)$, which we get by slicing the distribution at the value $Y = y$.



Question: What would this plot look like if X and Y were independent?

Linear Gaussian Bayes Rule: You can find the conditional distribution of 2 multivariate normals by using the law of total expectation (see notes from section 0) and multiplying together the PDFs, but you want to avoid doing that when possible. The formulas below give you a useful closed form solution for this in specific cases. You can reference them in Murphy 4.124 and 4.125.
Givens: Given the following:

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mu_x, \mathbf{\Sigma}_x)$$
$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{\Sigma}_y)$$

which says that x is drawn from a normal distribution parameterized by mean μ_x and covariance μ_x, and y is drawn from a second normal distribution the mean of which is a linear transformation of an x sampled from the first distribution ($\mathbf{A}\mathbf{x} + \mathbf{b}$ for some matrix $\mathbf{A}$ and some vector $\mathbf{b}$), and the covariance of which is $\mathbf{\Sigma}_y$. One of the names for this is a linear Gaussian system.
Posterior: We can compute the probability of x given y using the following formula:

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mu_{x|y}, \mathbf{\Sigma}_{x|y})$$
$$\mathbf{\Sigma}_{x|y}^{-1} = \mathbf{\Sigma}_x^{-1} + \mathbf{A}^T\mathbf{\Sigma}_y^{-1}\mathbf{A}$$
$$\mu_{x|y} = \mathbf{\Sigma}_{x|y}[\mathbf{A}^T\mathbf{\Sigma}_y^{-1}(\mathbf{y} - \mathbf{b}) + \mathbf{\Sigma}_x^{-1}\mu_x]$$

Note: Knowing when and where to reference this formula will save you a lot of algebra!

1.2 Exercise

Question: Compute the mean of x given y ($\mu_{x|y}$) given the following.

$$p(\mathbf{x}) \sim \mathcal{N}(\mu_x, \mathbf{I})$$
$$\epsilon \sim N(0, \sigma^2 I)$$
$$\mathbf{y} = \mathbf{A}\mathbf{x} + \epsilon$$

Solution:
Re-write: We start by re-writing this in the same form as the formula for the Linear Gaussian Bayes rule we just saw.

$$p(\mathbf{y}|\mathbf{x}) \sim \mathcal{N}(\mathbf{A}\mathbf{x}, \sigma^2\mathbf{I})$$

From this, we can see that $\mathbf{b} = 0$, $\mathbf{\Sigma}_x = I$, and $\mathbf{\Sigma}_y = \sigma^2\mathbf{I}$.
Plug into Formula: We plug it into the formula above and get:

$$\mu_{x|y} = \mathbf{\Sigma}_{x|y}[\mathbf{A}^T\mathbf{\Sigma}_y^{-1}(\mathbf{y} - \mathbf{b}) + \mathbf{\Sigma}_x^{-1}\mu_x]$$

We see that this depends on the covariance.
Solve for the covariance:

$$\mathbf{\Sigma}_{x|y}^{-1} = \mathbf{I}^{-1} + A^T(\sigma^2\mathbf{I})^{-1}\mathbf{A}$$
$$\mathbf{\Sigma}_{x|y}^{-1} = \mathbf{I} + \frac{1}{\sigma^2}\mathbf{A}^T\mathbf{A}$$

Solve for the mean: Plugging this back into the formula for the mean, we get:

$$\mu_{x|y} = (\mathbf{I} + \frac{1}{\sigma^2}\mathbf{A}^T\mathbf{A})^{-1}[\frac{1}{\sigma^2}\mathbf{A}^T\mathbf{y} + \mu_x]$$

Question: What is the relationship between this and ridge regression?
Question: How can you apply this formula to get $\mu_{x|y_1,y_2}$, i.e. the mean of x conditioned on 2 independent draws of y?

2 Linear Regression

Note: Wasn't planning on going through this in section unless there's a bunch of extra time.

2.1 Review

Model: We are given inputs $\mathbf{x}$, these are our data/features (we don't optimize these). We use these to "predict" outputs y, which we do by finding the mean of the conditional distribution $p(y|x)$. The prediction y is a linear combination of the elements of x, which is why we call this linear. We call this regression because the y values are continuous.
Inference: Solving for weights in linear regression corresponds to an orthogonal projection of the known outputs y_i onto the column space of $\mathbf{x}$, that is, the linear combination of $\mathbf{x}$'s features that reduce the distance to the true y. For our set of parameters θ:

$$p(\mathbf{y}|\mathbf{X}, \theta) = \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I})$$

where $\mathbf{w}^\top\mathbf{x}$ is the linear term and σ^2 observation noise that is considered to be known (you can learn it, but it requires some extra work).
The log-likelihood for $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$:

$$\log p(\mathcal{D}|\theta) = \sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i, \theta) = -\big[\sum_{i=1}^{N}(y_i - \mathbf{w}^\top\mathbf{x}_i)^2 + \log(const)\big]$$

Maximum Likelihood Estimate for the weights ($\mathbf{w}_{MLE}$):

$$\arg\max_w \big[\sum_{i=1}^{N}(y_i - \mathbf{w}^\top\mathbf{x}_i)^2\big] = \arg\max_w \big[(\mathbf{y} - \mathbf{X}\mathbf{w})^\top(\mathbf{y} - \mathbf{X}\mathbf{w})\big] = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

Bayesian Linear Regression: To make this model Bayesian, we now put a (normal) prior over the weights $\mathbf{w}$:

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$$

Our likelihood is the same as before:

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \sigma^2\mathbf{I}) = \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I})$$

The posterior over the weights is proportional to the product of the prior and the likelihood. Writing out this product and completing the square is hard, but remember that we have our closed form solutions for the parameters of the posterior of a linear-gaussian (see previous section for formula):

$$p(w|X, y) = \mathcal{N}(\mathbf{y}|\mathbf{m}_N, \mathbf{S}_N^{-1})$$

$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \mathbf{A}^\top\mathbf{\Sigma}_y^{-1}\mathbf{A}$$
$$\mathbf{m}_N = \mathbf{S}_N\big[\mathbf{S}_0^{-1}\mathbf{m}_0 + \mathbf{A}^\top\mathbf{\Sigma}_y^{-1}(\mathbf{y} - \mathbf{b})\big]$$

Comments about Bayesian Setup: The posterior is a distribution instead of a point estimate like the MLE. We can take the maximum of the posterior, called the MAP (maximum a posteriori estimate), which is a point estimate, or we can compute the posterior predictive which integrates out the parameters when making predictions by using the posterior distribution. The second option is what people are talking about when they say something is fully Bayesian (but you don't need to know that for this class).