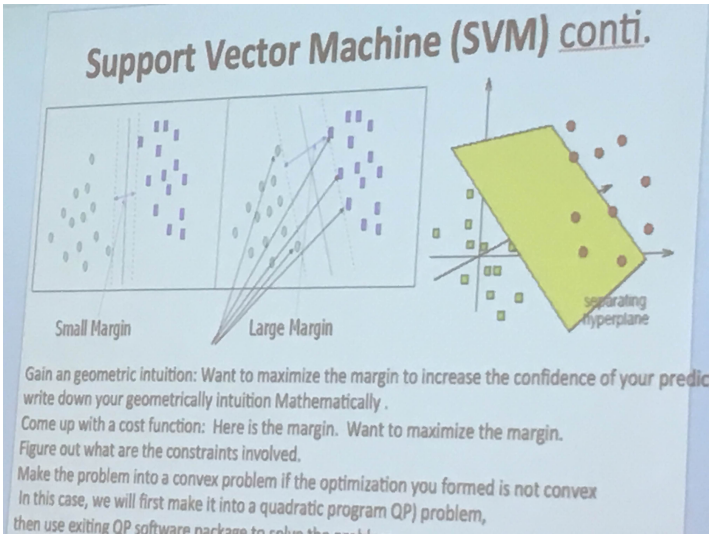
Mathematics of Big Data I

Lecture 5: SVM, L^1-Regularization, Naïve Bayes, More on the Gaussian

October 3rd, 2016

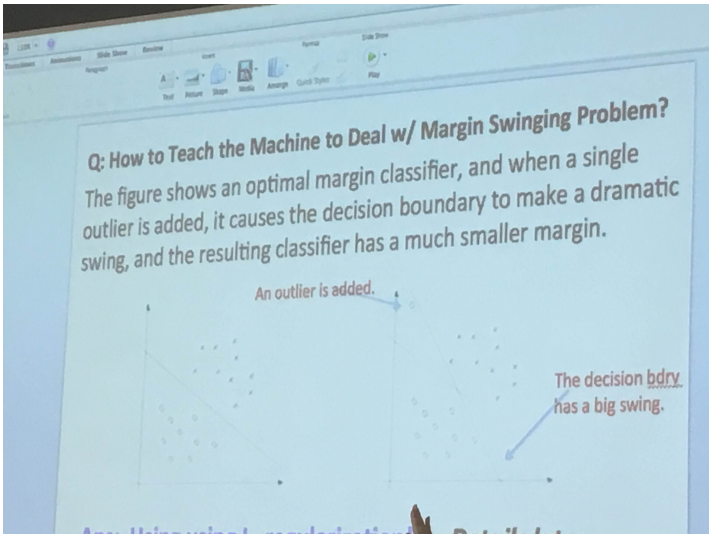
1 SVM Continued



Recall that the support vector machine is a means of classifying data into binary groups. We want to maximize the margin (distance between distinct classes of data) so we have greater confidence in our prediction. Generally we would like to formulate this in terms of a convex problem, but if it's not then we can formulate in terms of a quadratic programming (QP) problem. Another method is by using *kernel methods* which allows us to solve a larger class of nonconvex problems.

When we derive the support vector machine, we use our geometric intuition for assigning a cost function to our given problem. One question we have is the following: **How can we teach a machine to deal with a margin swinging problem?** What this problem is that the decision boundary is very sensitive to outliers, which makes the decision boundary "swing" dramatically with the addition of new data points.

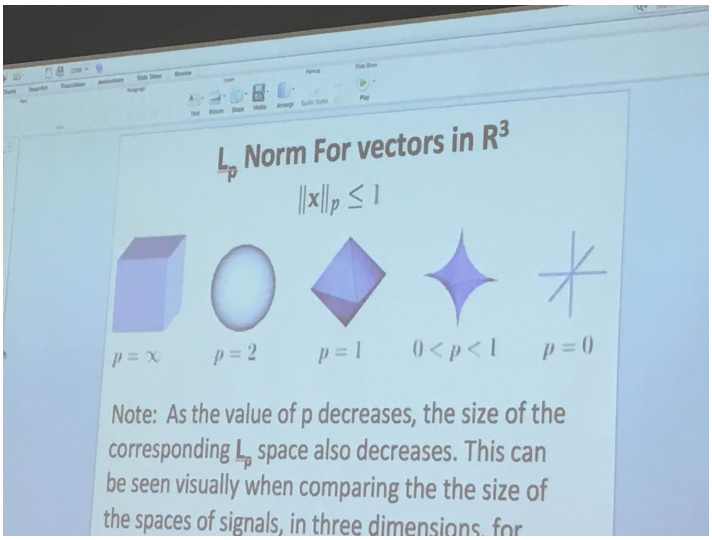
2 Regularization



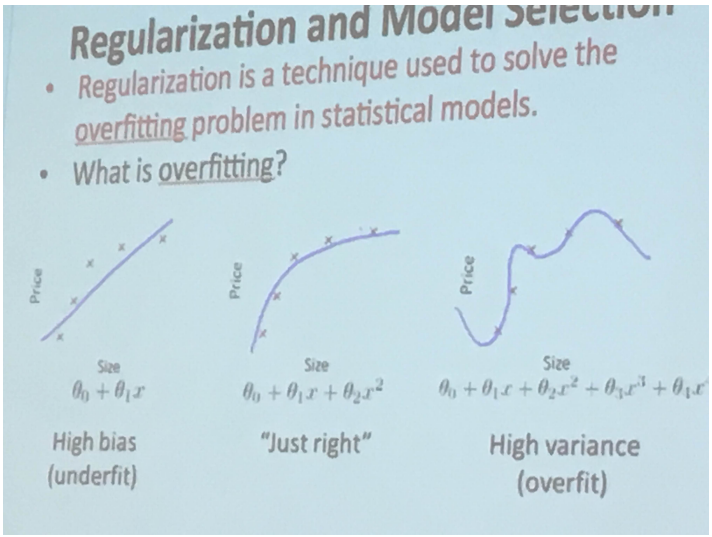
The answer to this is L^1-regularization. Recall that the L^p norm of a vector $x \in \mathbb{R}^n$ is given by

$$||x||_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}.$$

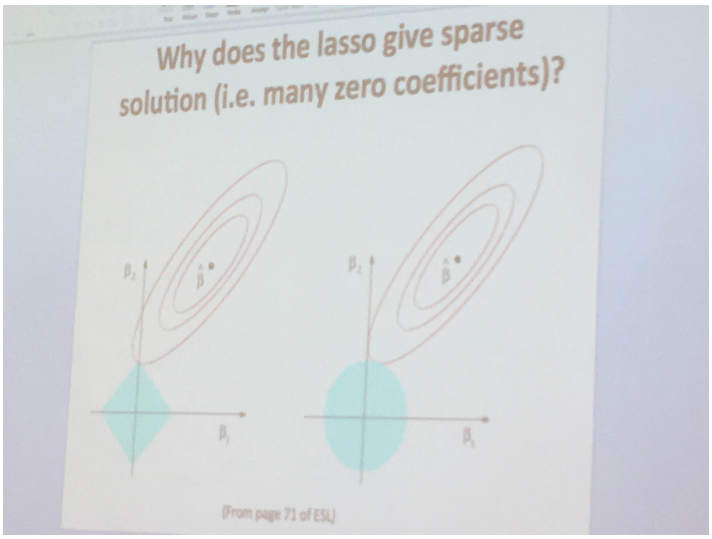
We see that these are different metrics and therefore, viewing $\mathbb{R}^n$ as a metric space with respect to these different metrics, we yield different models for open unit balls. For instance when $p = 2$ this yields the standard Euclidean norm and the model for unit ball is the solid sphere (a disc when $n = 2$, a solid sphere when $n = 3$, and their higher-dimensional analogs).



Regularization is a technique used to solve an overfitting problem in statistical models. Overfitting happens for example when we fit a polynomial to a set of data point in $\mathbb{R}^2$, but our fitted curve passes through every point of data without capturing the macroscopic structure of the data.



One method is called a **lasso**, standing for **least absolute shrinkage and selection operator**.



Here we consider the L^p-ball around the origin in our parameter space, and draw ellipsoids around our parameter point. We expand these ellipsoids (maintaining the same shape) until we intersect the unit p-ball. The nearest axis to which the ellipsoid intersect the ball tells us which variable among our parameters is the most important in our model. This is why we are able to obtian sparse solutions with the lasso method.

3 Bayesian Approach

Many times in data analysis we use a geometric approach, say like support vector machine. However, another approach to data analysis is to use a Bayesian approach, which utilizes techniques from probability to solve a problem. Here, no model is needed for the problem but everything simply uses methods from probability. Within Bayesian methods we can have **naïve Bayes models** and **naïve Bayes classifiers**. If we have a probabilistic model the "naïve" part is to assume that multiple observations are IID random variables. Recall that

$$p(C_k|x) = \frac{p(C_k)p(x|C_k)}{p(x)}$$

where $p(C_k)$ is called the **prior**, and $p(C_k|x)$ is the **posterior**, $p(x|C_k)$ is the **likelihood**, and $p(x)$ is the **evidence**. In naïve Bayes we assume that each feature is conditionally independent of every other feature. We have a chain rule in probability such that if our probability we're looking for is $p(C_k, x_1, \ldots, x_n)$ we can decompose it as

$$P(C_k|x_1, \ldots, x_n) \propto p(C_k) \prod_{i=1}^{n} p(x_i|C_k).$$

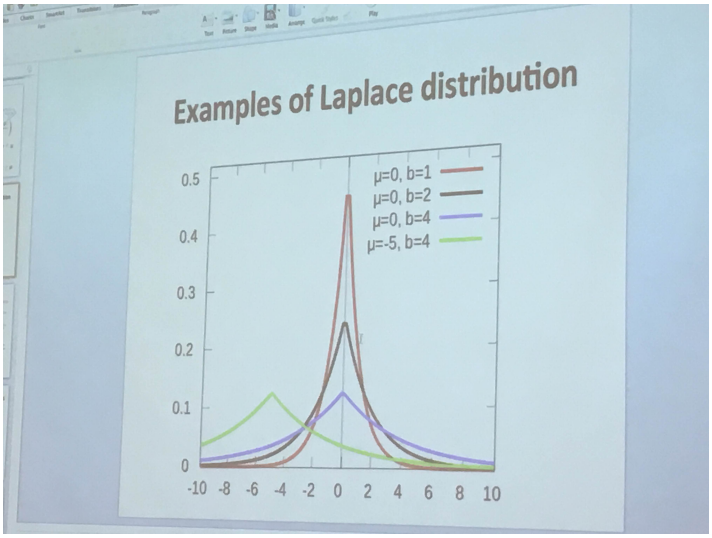
A **probabilistic classifier** is a classifier that is able to predict a probability distribution over a set of classes, rather than only outputting the most likely class that a sample should belong to. Probabilistic classifiers provide classification with a degree of certainty. Binary probabilistic classifier are also called *binomial regression models* in statistics. Oftentimes what we are trying to do is solve an optimization problem

$$\hat{y} = \arg\max_y Pr(Y = y|X).$$

We want to illustrate how this may work with the Laplace distribution, where $x \sim Lap(\mu, b)$ if it's density is

$$f(x|\mu, b) = \frac{1}{2b} \exp\left(-\frac{|x - \mu|}{b}\right).$$

Observe how the L^1 norm is a part of this.



Another important distribution is a student t-distribution, which mimics a Gaussian but has the advantage of having compact support (whereas the Gaussian has the entire real line as it's domain.

4 Review Topics for Midterm

4.1 Ridge Regression and Lasso Method

Recall that ridge regression is simply an L^2-regularization penalty added to our cost function. Note that lasso is a regression method that involves penalizing the absolute size of the regression coefficients. Penalization forces some of the entries to be exactly zero which encourages sparsity in our solutions. This is useful when we want some automatic feature selection, or if we're dealing with highly correlated predictors, standard regression may yield coefficients that are too large. Here we can compare our regression problems:

$$\begin{aligned}\hat{\beta}^{ridge} &= \arg\min_{\beta\in\mathbb{R}^p} ||y - X\beta||_2^2 + \lambda||\beta||_2^2 \\ \hat{\beta}^{lasso} &= \arg\min_{\beta\in\mathbb{R}^p} ||y - X\beta||_2^2 + \lambda||\beta||_1.\end{aligned}$$

While the models are quite similar they can yield vastly different solutions to their respective optimization problems. The parameter λ controls the strength of the penalty. We want to balance two competing ideas in our model: fitting a linear model y on X and shrinking the coefficients. The result however the L^1 regularization term in lasso regression is that the model will force some entries in our parameters to be exactly zero. Increasing λ suppresses coefficients to be zero and there is more shrinkage.

4.2 Marginal and Conditional Probabilities of the Multivariate Normal Distribution

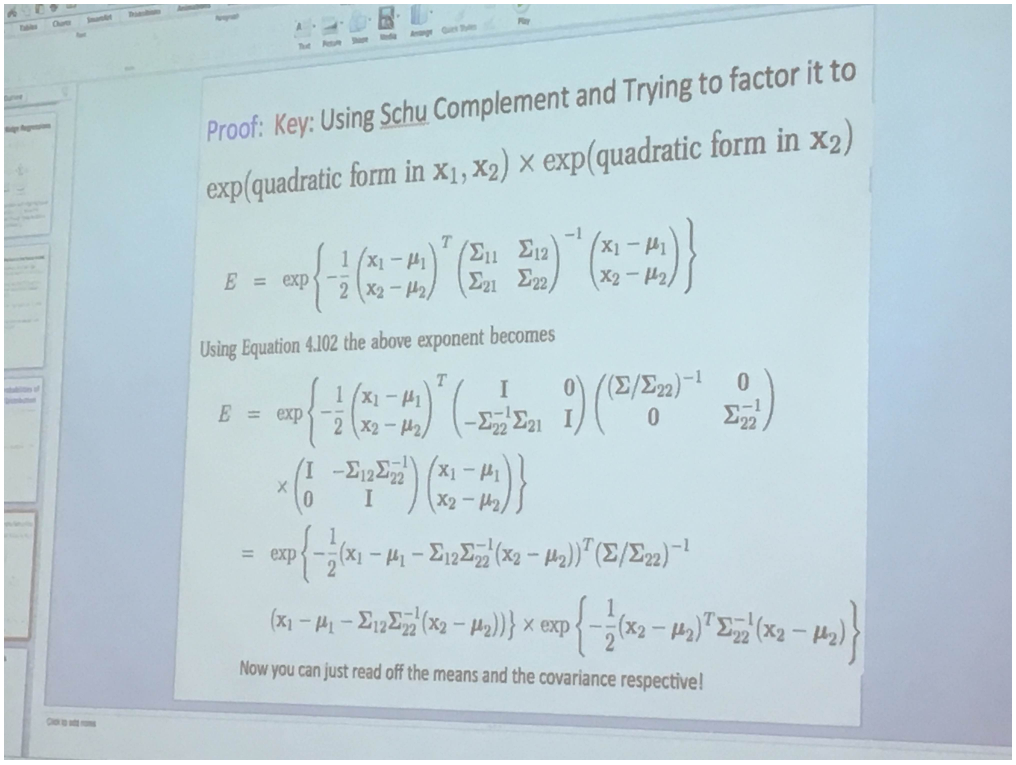
If we're given jointly Gaussian random variables $\mathbf{x} = [\mathbf{x}_1 \ \ \mathbf{x}_2]$ with parameters

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \qquad = \Lambda = \Sigma^{-1} = \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix}$$

then the marginal distributions are given by $p(\mathbf{x}_1) = \mathcal{N}(\mathbf{x}_1|\boldsymbol{\mu}_1, \Sigma_{11})$ and $p(\mathbf{x}_2) = \mathcal{N}(\mathbf{x}_2|\boldsymbol{\mu}_2, \Sigma_{22})$. The posterior conditional is given by

$$\begin{aligned} p(\mathbf{x}_1|\mathbf{x}_2) &= \mathcal{N}(\mathbf{x}_1|\boldsymbol{\mu}_{1|2}, \Sigma_{1|2}) \\ \boldsymbol{\mu}_{1|2} &= \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \\ \Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \Lambda_{11}^{-1}. \end{aligned}$$

The proof of the above decomposition is given using the Schur decomposition detailed in the picture below:



To discuss principle component anaylsis, take $\vec{x}$ to be a vector and let it's covariance matrix be given by $\Sigma = Cov(\vec{x})$. We know that since Σ is symmetric positive semi-definite, then there exists a matrix P satisfying $PP^T = P^TP = I$, a diagonal matrix D of the corresponding eigenvalues and $\Sigma = PDP^T$. We can rearrange these matrices so that the eigenvalues are in descending order with the largest eigenvalue in the upper left corner of D. In general if we have a data matrix X which is $d \times n$ in size we can find the diagonalizations of $X^TX = VD_1V^T$ and $XX^T = UD_2U^T$. Note that while D_1 and D_2 are different sizes, they have the same nonzero eigenvalues. This gives us the singular value decomposition $X = UDV^T$.