Boosting a decision stump

In this homework you will implement your own boosting module.

Brace yourselves! This is going to be a fun and challenging assignment.

- Use SFrames to do some feature engineering.
- Train a boosted ensemble of decision-trees (gradient boosted trees) on the lending club dataset.
- Predict whether a loan will default along with prediction probabilities (on a validation set).
- Evaluate the trained model and compare it with a baseline.
- Find the most positive and negative loans using the learned model.
- Explore how the number of trees influences classification performance.

If you are doing the assignment with IPython Notebook

An IPython Notebook has been provided below to you for this assignment. This notebook contains the instructions, quiz questions and partially-completed code for you to use as well as some cells to test your code.

What you need to download

If you are using GraphLab Create:

- Download the Lending club data In SFrame format: lending-club-data.gl.zip (https://eventing.coursera.org/api/redirectStrict/Xw4aYrEBNia6Mq9R6-MNhKpxPZ2t-sN5nCkbfYSNgDg0rX6DxDTBnml8J2b5auAJPxnJMGJBY-fbRFbQBIOIUg.mAbK5dUoUHBPwKrVf-4AaA.bXYHAIGp18huRmeAsXhnaXRKTIKF71dGaAYC3vSAWTaTiL1aTOz4rwIObL0jTU8FIJST82meVATnm6fHeHOM1v-i_SnK1paq7UPjfAPS96trJRpob_iW72GOSPopTwUGipIOg_na3UoxnBmgVlVi9hT-0vYcDE9cnttkfTskYYe2qmb9-5c2M2X1rcQlH4EFwZzLR0ms6Oe_Btwxt45tQvh4OeRpXwngxsduNzjdpMWZCKx01D28vKCLSMQ0b9_7lMCKTY2RxWkMGgQiWa9kkNHEdSfLlxgbcP2txBkPRhbIaiwtoDQH5bs6UTDKWLZ7EKxdsOYQ65bWGgd6VOl0IBSua9N_xNsJGTdnst5LeNtBc2_fIEoEJjgZHOACLERei7hKO61QnqbPj65krRZZ7buw9P4YCYxC1GT0YcItsuxHwnAW8bgLboR8dHiPfMtE)
- Download the companion IPython Notebook: module-8-boosting-assignment-2-blank.ipynb (https://eventing.coursera.org/api/redirectStrict/tqTgeFmJcHMs5U_ncwrQVOEs-_f79qdvQT-iK2zOSu1b-QC4mR2Yz4FyysFOHWgL1eFsbEjmfw1tqz2T7xmgdQ.fwOXDVkQgGBRRDydHLSZYw.gmC-3lklhL-DSvBcxnjmga6YLwvl4e1JV-WX7NGD8y_XaaKBoLfXf9pKGVbPMrF1qgxE7pOm-pNmtvY7T-4gRPyAYk2EfubEXZn7T_3heSot1x9Dr1x0ZOJ3R2_eiTFGLcJ64uhGs6qpWScuEA25afEjT-5Vgc7jOaUP0xM3J43clqAdjBn5gcYhHNE91JCR1TYeEBH7dCHMH2wBBmRo3G5FTNsZA3uB2ruWbz5t8HIbLnlyR7nU3qb0x_WQqmRkd960OWP4HjftX0wifZM8LCDulJWaJU6G-vV9pDy7wiXf1O51qdsy7Ys5zySL0qTzwxguBC-JBUiZJfIxaieKfatKhabKTxITx3UrDiwE54z3LhkrLvgc1IAYV2Q8kBgqWX9qUt_wQzEirkfPpqrCRMrqejHvBGEO8C8wdVdHmU5D7DkrZiKqPE3-WtqjRVtf50LCuRjw3WK79oj438MmuE3YDeSeEe_HtmkRbTPSU4IEKmi4QWuCR588dnJAHvIC)
- Save both of these files in the same directory (where you are calling IPython notebook from) and unzip the data file.
- Follow the instructions contained in the IPython notebook.

If you are not using GraphLab Create

- If you are using SFrame, download the LendingClub dataset in SFrame format: lending-club-data.gl.zip (https://eventing.coursera.org/api/redirectStrict/Xw4aYrEBNia6Mq9R6-MNhKpxPZ2t-sN5nCkbfYSNgDg0rX6DxDTBnml8J2b5auAJPxnJMGJBY-fbRFbQBIOIUg.mAbK5dUoUHBPwKrVf-4AaA.bXYHAIGp18huRmeAsXhnaXRKTIKF71dGaAYC3vSAWTaTiL1aTOz4rwIObL0jTU8FIJST82meVATnm6fHeHOM1v-i_SnK1paq7UPjfAPS96trJRpob_iW72GOSPopTwUGipIOg_na3UoxnBmgVlVi9hT-0vYcDE9cnttkfTskYYe2qmb9-5c2M2X1rcQlH4EFwZzLR0ms6Oe_Btwxt45tQvh4OeRpXwngxsduNzjdpMWZCKx01D28vKCLSMQ0b9_7lMCKTY2RxWkMGgQiWa9kkNHEdSfLlxgbcP2txBkPRhbIaiwtoDQH5bs6UTDKWLZ7EKxdsOYQ65bWGgd6VOl0IBSua9N_xNsJGTdnst5LeNtBc2_fIEoEJjgZHOACLERei7hKO61QnqbPj65krRZZ7buw9P4YCYxC1GT0YcItsuxHwnAW8bgLboR8dHiPfMtE)
- If you are using a different package, download the LendingClub dataset in CSV format: lending-club-data.csv.zip (https://eventing.coursera.org/api/redirectStrict/sUAUayVmlaSvm8XmuYtw8JCVj4xwJWBC9Dw4krY4L4hQzY6bqJ21XN4_Rb-6ypI3AaL_a0lORrL3PPvbkY52xA.kfIIC732nujFiHAsKB2mow.x2o4oHKJz7GEO3rPaSHYzuDps5gtr0EU4HUpw6bECo21iLFH15Z-4amVpHjf9E1xjUgtU673uuPi5Lr07lpJZft0qf-SSPdzsuIZkqlUpg7SYY6b1RjK_cwn74He-uPrv59pDkYniTSptk7XyYQb4o9tfWVMyk7eZynXudLfvdWzgzrhEb_FWpapKT7kWSAnIRTHwH3RGfQYEkmuXsxm2GI94_Zn4fhJB8ruMTFN9FiJPBSA0cV0CedWF2GwUsHbnEV1q8KXpWiw86Zjmk6vbVvFiNKqzCW-W9yxQMaL4geBSmYzIrW-JfRbz63RrqfjetV2seF3B-rJsz7oEPg-KhZ2YKHj-4nldf7uPGg5oO_TlHXhay5t26QmkYHWraQ9MTth_Z0X9KrifT7M00C9CeYjpha5xiPYflQNnXO42FgK83KWoGJH-iWPVJpDYHau)

If you are using GraphLab Create and the companion IPython Notebook

Open the companion IPython notebook and follow the instructions in the notebook.

If you are using other tools

This section is designed for people using tools other than GraphLab Create. **You will not need any machine learning packages** since we will be implementing decision trees from scratch. **We highly suggest you use SFrame (https://eventing.coursera.org/api/redirectStrict/GANIQjChfQ0lqK1W68fn8U9TeKhFSAmruFGMWNsRtEFD2Vb0Pm6pDj39XnXAhlDUgwFuVxXsIpOOk-ZoEar3CA.EaFN2HDItXBewbpXozz6mg.JBSYkqrdEKXSiZfLB_gnDWr07cMNSzvl9dvntGz9REuDyjB_Il2ZRgt28L2wYBR-_91OwXaNcO3Q-rvlWOCxkdNVVzPqHXXkORZviRkAuS23C2GU6-1sXnQUN5d3LGwn59dyID88oB-DNdxfdTk5t-oS7cUdvnW2x4fH7TrHAQrCXIIcStwenqmW46loMqedzxyoS-uOplcYOmGt6p_BaS6Ai4jVZybg_wC-EUYA59GbqiVkoKKr-_VBu1ca9OHCBCxvHaoM6NFClhVZGb-pF99j3GtVyylmRHiESu7DXV6TbWA325cxzoZrOSBI_W6-) since it is open source.** In this part of the assignment, we describe general instructions, however we will tailor the instructions for SFrame.

- If you choose to use SFrame, you should be able to follow the instructions in the next section and complete the assessment. **All code samples given here will be applicable to SFrame**.
- You are free to experiment with any tool of your choice, but **some many not produce correct numbers for the quiz questions.**

If you are using SFrame (https://eventing.coursera.org/api/redirectStrict/GANIQjChfQ0lqK1W68fn8U9TeKhFSAmruFGMWNsRtEFD2Vb0Pm6pDj39XnXAhlDUgwFuVxXsIpOOk-ZoEar3CA.EaFN2HDItXBewbpXozz6mg.JBSYkqrdEKXSiZfLB_gnDWr07cMNSzvl9dvntGz9REuDyjB_Il2ZRgt28L2wYBR-_91OwXaNcO3Q-rvlWOCxkdNVVzPqHXXkORZviRkAuS23C2GU6-1sXnQUN5d3LGwn59dyID88oB-DNdxfdTk5t-oS7cUdvnW2x4fH7TrHAQrCXIIcStwenqmW46loMqedzxyoS-

uOplcYOmGt6p_BaS6Ai4jVZybg_wC-EUYA59GbqiVkoKKr-_VBu1ca9OHCBCxvHaoM6NFClhVZGb-pF99j3GtVyylmRHiESu7DXV6TbWA325cxzoZrOSBI_W6-)

Make sure to download the companion IPython notebook: module-8-boosting-assignment-2-blank.ipynb (https://eventing.coursera.org/api/redirectStrict/tqTgeFmJcHMs5U_ncwrQVOEs-_f79qdvQT-iK2zOSu1b-QC4mR2Yz4FyysFOHWgL1eFsbEjmfw1tqz2T7xmgdQ.fwOXDVkQgGBRRDydHLSZYw.gmC-3lklhL-DSvBcxnjmga6YLwvl4e1JV-WX7NGD8y_XaaKBoLfXf9pKGVbPMrF1qgxE7pOm-pNmtvY7T-4gRPyAYk2EfubEXZn7T_3heSot1x9Dr1x0ZOJ3R2_eiTFGLcJ64uhGs6qpWScuEA25afEjT-5Vgc7jOaUP0xM3J43clqAdjBn5gcYhHNE91JCR1TYeEBH7dCHMH2wBBmRo3G5FTNsZA3uB2ruWbz5t8HIbLnlyR7nU3qb0x_WQqmRkd960OWP4HjftX0wifZM8LCDulJWaJU6G-vV9pDy7wiXf1O51qdsy7Ys5zySL0qTzwxguBC-JBUiZJfIxaieKfatKhabKTxITx3UrDiwE54z3LhkrLvgc1IAYV2Q8kBgqWX9qUt_wQzEirkfPpqrCRMrqejHvBGEO8C8wdVdHmU5D7DkrZiKqPE3-WtqjRVtf50LCuRjw3WK79oj438MmuE3YDeSeEe_HtmkRbTPSU4IEKmi4QWuCR588dnJAHvIC). You will be able to follow along exactly **if you replace the first two lines of code with these two lines:**

```
import sframe
loans = sframe.SFrame('lending-club-data.gl/')
```

After running this, **you can follow the rest of the IPython notebook and disregard the rest of this reading.**

Note: To install SFrame (without installing GraphLab Create), run

```
pip install sframe
```

If you are NOT using SFrame

Getting the data ready

We will be using a dataset from the LendingClub (https://eventing.coursera.org/api/redirectStrict/MmtyZbeYlR7QeQGk_JmfrQXi9L6Co8MVEkUuVkg49n3_Ac-wjE6FofxCFZ078bp2vEah-njuF_OJ3N6w6EHi3g.V9_RyKetaUxRL0t5H92pKw.5wPIn9ePfkKO5D2P0xlkAWINK5aQ4aInBKamAGHAeLF3E5W5_hiwIS157-sjpC1_xnmSlsVUMlAS6p_JWZytgh1jhqSLHo6AmvBShy4k18gSpniyRpr95_XqJxp_wnLntvKUDPBCwJXUJ5OobityWGHlzc9ZVDo3mDX1dGjzBuAQiGdJ6Aj5P3ix0U21sFp_urvZUkqjY_ejTFeBQvI4P3-Qk7qXNeaUC4fFunsbd2oKq2GdDUJRIFIM6k3AVi1C_mniaDi7s__oDMNjlk6KXRHdp-ZpqJbVbB4Szeo8Ewt72cUva5fo8tdiWPSQC0YFih8v13L_nB5QOJTPeSAFoNM0vgc8SfvdvaF2cQD2kcB_apbJxSjPXn-uwvhNqGJN7XdISjE5FsA9iVrhEcAh9KEF9WsYyqYW8Re-sksgobyvkmgYyKv2F1CnrxdautbuA0Ibgr1duVHnUXfCtv-vysPDM1H5w5rD7yp_mkPIGIeZcZNgiWhyr0BU5_qDPzw7D6AZxZ-8vEIVUSYmlhqn8Tzr_fUKEHTmkqQ3iEx9qH-F9oseWuQqV5dP1SEpgSa6laur2sJ0MP3jneL4-H18aarx0WMm9-0d50zyj286ZXO-5P858nqmliAtavtUhMNg3eJlqmIQPZ5mfALda2SZ-YWZWxpOYo3py8vZU1TMSQIlCue482Ca2HXuyQb9T2DhzddbhgzhOMET6e2pqSCZjz0GaKNvfS-oY5Qmo1PMtzA5bN6UVRkh_v8R7rWCR6lXDkNEb31x7WZL3bt1GMGweuvPYdUBzqPKAnBsi_hNlz6ehr-D9V-tXgwlhAV_O80V8OVU6Ex_8AmF84O6IHX1XqT9xIETivb0__V_Va9Zgu5kJwBBhINX3xtitMO9_6QU3pwKFnVAPZicEeFUUSJB3ApUFgWKbFV35oobDyzfl6bQy1nkO6qUptozMFlhIG0sSMnSuh5j-1ihNIKQVu--wOb0l1uZiQatpZ-czNaYHlUPhaxfXYVHpwkCd5BToUSgYNbD4_ogXr5s87HoDvqVz5m9cNMCPsWnWkY31muQ3Zeh6ERw9fvAzGqJv_m6kiL6JT5rJGv2L134WB37MxzwQ_yHxTo2xJAQfVfy1FvDmKGInqPNy3gvDDmIxoYBl5GriFbEXzfXWOvU5Pp2OYsOXmvsaKP0Hfjm-N23T8RlOqq0S-jWD6LWbvMUY2OdgP2iw8ahyTUugYXPuW_kAGIF1RO0qN16tFR6_LrBMn-Zgsa8hu12Lx10EqkZUPUPk9uGpGRtQGaF7Oe5W8hMUmb-fWkuL70vz2urF9Kobe3dODFEQK6xNSCradvMzZaRqKKeFU5g_k6bTqu_b7BOAFKWGu-Uh1Ed__qFsFo34zjvbbZe6QUYzc0NzDG3-xhrLa_nMfQeS-XWQce4IPnV-4qyF1orB37qIKyq1yin22nujY1yzx1Pzwy1nOC8VMpSW36tbaCRPvKom7kT3Fv-iLp2pBDWAK7Iks_Ebu9Kt3B3bS55pYCHiBqdSdiQCfgsjsM0MiHsH2d-Y-hr_lxc9dET6ptPBDEjuFJW6xCX7CdX7nFPLRUJcZvO8UAxt8n-

w9ZV9aQIcQuWdLlb8X97QH3KLvArjzrYIJFpR9AfqB4iHRt4da8DHtNMGJvso68kfcn47C-F9oFgE16xQGaLoxt5FxdtlxekOaDE56qCQW-sDtG_dasoUoZn3LmASGiXfgI0zkQei0-VZU5KLR8N9gXg0flHRnHc9PI4KAmtVs5lPaKiQ4nxGCSFlakXmpUipgp0kHMYXTbWZHvGZNv3wKwMYuuKGpjPzSg2jc8R4V-Dp_ITrKrKxEXBfdu-jpbLp6lC9zVFqFLrrmxULAijFMvOHrwvRvlRImT64vNY-PmHE4TJmhhPeU_V2hAVcIKCHMloDa26rEKDTAMBeuvyjFUZZmt-IGpuV3Oy9oe7FJvNeangXSlvkKWBTeNNdImT8SCbkNWdQa_VYq3wfVeTwduVMmRKTTukbyUxIpo_Gs5lSJDpGX_aUSrCsWU52K0_OYdvl7zUMPIzgOcnEmnTlGo2__EQccD42Wf4E7cC4MqtJi34GjbsYyD_a1kcDxk2wwB1GARNdYJx_eJfkIypR3PectBnUasenFf-4x14Yql8PCutCANy86g5Cr_F1OAiW6J_4p6DIbUxeq62clqmRejTHaFcYQ5ggBjU-PCEkG9yAXGM6NpOUCh79hcIPyrHsuXC2u7FI2bMIkPydSeQGDlP3g_uEv8EpoLRSLpSScwm8ufy-mHLyhdTgUTTw5kVtETdun8vzFETNwngGyEURGTuX3-hdxXEyHKkjkIiOVBPkV8c3s1LarzOebBHBI0h_U9jWyJiNFRVPb0R4cfTvwpEGObIVx01FTPgycAgsL1vW2dAlwWdxXRw_Oja8jP65DF5dn72DqSJ6EyKwFezn-m5F2QuzGIFTFCQOX7gSp2VlT-4XcwbBBaXGG96c-RMVoJEvS65ibWzbJ8cBeTTDuTVtkFlY0gzKuVLLQku7sSsr-StnDuktX1923VYCkBMEAka6CyPYQ4dTD_Htl059far9k770i0oqNKzYy2q0gWRNH493CKOatNz5NmsAvAcwZH9jDhMcxpzdwTD0FonB_OICISA70jBkpy0_E4IP-oGMoC5ys_3h7vwJO3q0MJW6QqQkv2LWb3_PN5ZxcyBRZR0atQYuxqGnkaxgFCh5P2BOWt6I5S42pKaPQYgEgnkf1ufAqlVnLEmGC2YpAf9BTVzaMs8HOO1s_FCn30uws_BREgkwPxg3aw3gMW7LODF7bWp2BaER3FJ8qacTgmzLg02MbBouHNIaNgwDQOfbVeA--8SnKShOyDxS-vAv1HhHtfzbZZtVUHzstnnjoRMUbWxEEqKuGJedgatvyq8uThmA7M7VKLkuWuj4XYxrov7TsQFl99b4PhfW9jMZnCx67kNpPJlgmbQdqcoPpdfm3m_PfTLFpV9hZV7GgpPpXHLejI50yldLl1mXk0NXnBnvidbUIxOvIbMgdnBpUjRfLzIjtnPApOZdx3JKDONPf9oWeLBMY360EAyfcJ_PBlFyVysRfJwEhg2sU9L1z5eRMNaUbxiVJAwhZj7QccVKeG0tk9DK6hj4DsN9VahFBj-SeD5xJF0MCZy6LvLQHvCW_3QHUNRCesYUTCJUXboJEj4JeMW8YqY5B8iRZuuBhAFX-bs5o6L_PSXdwHTYAsvwzIhpkwSNEBKn1jaPS1GAtGX5tZAS8WK9ZHRFmS2F28SeqSMdtG_MLAFOIYSgwkndtzDa4MAiwcsmvZbmehVIlvq7TRsVbZh1xX4rDx1K6oK3u15Ofo3k1MDIPgj3jC7aACYLEBB4tAGm0_VGdfqI89_Od_c4xXJ2n-4AWrZquYLgJ0DfcIPP8Dro9_JJyW6bau4LcZ1Xg7ps1cL1usqvu299Js90LkGTcRG-huC01NaGiOmOY2pKpIGyeRzFpQYfQH0T8wpJBk4omFMYjVXYEf8gpgDNECDzD8wTeN1S4UwX4Pu8wHpvmcB-wqebRJOO9UdxdnXzFvrYbsx8c6XVZBr2TJ0M9B-fJbxoeibwp8tEetEoY2l7bhJUbGq9Tmre47KxuVdzt6ovF3MgTfO0C9HJZi9pbUbMyGxv89HCq9NyLBmz4oV9pggORYO6jYB5qdDnp_w0cywvu3ndRMx79U2b7ydLkbjI0hCtrxIRI1_8FDMtZykdV9DfQ2lLgJ0_jYv7ia-0Cz1fniTqP4rRuaMo579eSowx6ntzT0MWa-9-OqKR9Qr8LdKGqEYQbQYNceEVE383UyDPmz_OpcdBf4B0Y_LpdA_VhV6vlE3cTUTqQ-I6GRwkdiVZ6xNTyV4Bf4VN5907yVnYN5Nyl8x_A2bW5dz5YEskYqIDLbaHSPeJ3jySeQhe4_T2_xL9KdkgfwvTdaiS0Q8iAN6kv9k6EEfmzAIVRGTeFUgJ18U5nV9yrdDshvhHK-ui3Iy6-sp6gQfJ6osXuYzbIIxy7cv5WHmdDyLR7Ild8oX_hVbha1JFlAWzBcjqGHHbmD8sx542RwuJ3f8WFWVHTeAy-MbloG2zMUC57eel2p-5mQFr-nMngfyH2Mgi4FCCPDHpA4TSnH1fcqlE0Z4F8STUHibI3HIgNxwh7CKeV-G_HTxtxSFe9ujOHW7E2Uf4qTTUtmrU-Gy4uAYbnupnh5vfp0n4orrArwlqS5p9r_EBODIZMeyhuFjXZyxdeWlT5zpFdvuxYHGWWY7D3TKCH7DSSEeWxnXXNI9-7tH_-q4iZrUonKmxmZ5x0kHrLPGP6TzI4wD8qdxu3cIQUFglxCvBWmJVxkuJGFKkYvAKg71B6mmqK-RBplJaBiMhM3U-CWodqgbj1KRxIWj5042yaP6_pyL0h-LDzlFUc4f0gBocQIRD4kSIDLZmrSUB-uKxDEZnz93k9GkeyxggNIhlPeKvrNpklIWHXz-mUjYBNYwqj39wqaO--vAUw).

1. Load the dataset into a data frame named **loans**.

Extracting the target and the feature columns

2. We will now repeat some of the feature processing steps that we saw in the previous assignment:

First, we re-assign the target to have +1 as a safe (good) loan, and -1 as a risky (bad) loan.

Next, we select four categorical features:

- grade of the loan
- the length of the loan term
- the home ownership status: own, mortgage, rent
- number of years of employment.

Your code should be analogous to the following:

```
features = ['grade',              # grade of the loan
            'term',               # the term of the loan
            'home_ownership',     # home ownership status: own, mortgage or rent
            'emp_length',         # number of years of employment
           ]
loans['safe_loans'] = loans['bad_loans'].apply(lambda x : +1 if x==0 else -1)
loans.remove_column('bad_loans')
target = 'safe_loans'
loans = loans[features + [target]]
```

Subsample dataset to make sure classes are balanced

3. Just as we did in the previous assignment, we will undersample the larger class (safe loans) in order to balance out our dataset. This means we are throwing away many data points. We use seed=1 so everyone gets the same results. Your code should be analogous to the following:

```
safe_loans_raw = loans[loans[target] == 1]
risky_loans_raw = loans[loans[target] == -1]
# Undersample the safe loans.
percentage = len(risky_loans_raw)/float(len(safe_loans_raw))
risky_loans = risky_loans_raw
safe_loans = safe_loans_raw.sample(percentage, seed=1)
loans_data = risky_loans_raw.append(safe_loans)
```

Note: There are many approaches for dealing with imbalanced data, including some where we modify the learning algorithm. These approaches are beyond the scope of this course, but some of them are reviewed in this paper (https://eventing.coursera.org/api/redirectStrict/S5M3twuz-_pUOmexx4HllGLKguJYlCzeh71r3VALsBog3G-Z9oq3tuqGjyTbd8q6_ibdkS8RDVZ4QWXd9jgwqQ.KjeSAB7m-mePKDq2KOHUQQ.ctjxo5rAYY8tdf8W7M2U7LxOzW5-T9C6WhcnOFzSSfp-OswB3VlOstS4Vl2eIpSUNnhP2GpZuV_g7wbpAQOXKPhumAkNge7Il3Gq3tjpZb179wqEXJAY8aT7T7srnqPc99x5gtMIRKrdZC8TnYqC9fHa62qgadzkH4fCfHeeSTNpy7zycw4twkFiYWIG8RcNktNvHGanC_ArmFe6LVRfWYip3fRiXEpNLsdgIN11Do1LRKObCxQmNE4omImpssM9TDR_YgIoKC1RriSsgk7pgV1li-Q1IPca2OgcxFlGBTDVgNVr22OMWNRDMzzhRUJO2HIPGXjaZbNCR5P1etL36KrG3V5jdwLJlUWUe-C48ILEyZLV1H6FrKlh_tvYmy09EGt2_ST1touqZcVnsoGa8HSEAZicT9rb2ITNajIwWtW1wK024fSSNI9LIyyzDrP-_BReb9zacFmC1ENU_0gC7X3SXLm0lkIRxkjJiL9NLMhZN99AfpWg4qNB7spmUa5-zMnEipxYPUwu91FyfOg6jHJe8QRgg6TpSI0LZ41wtfc8ivlVnzzmNOOeG53nFozV0sesDtQGHZDVvjbc8WFrIKvfkNfdFpT3_-b_By6-3-FhvUBsAquV00E_w3X5BpBT). For this assignment, we use the simplest possible approach, where we subsample the overly represented class to get a more balanced dataset. In general, and especially when the data is highly imbalanced, we recommend using more advanced methods.

Transform categorical data into binary features

4. Just like the previous assignment, we will implement **binary decision trees**. Since all of our features are currently categorical features, we want to turn them into binary features. Here is a reminder of what one-hot encoding is.

For instance, the **home_ownership** feature represents the home ownership status of the loanee, which is either own, mortgage or rent. For example, if a data point has the feature

```
{'home_ownership': 'RENT'}
```

we want to turn this into three features:

```
{
  'home_ownership = OWN'      : 0,
  'home_ownership = MORTGAGE' : 0,
  'home_ownership = RENT'     : 1
}
```

5. This technique of turning categorical variables into binary variables is called one-hot encoding. Using the software of your choice, perform one-hot encoding on the four features described above. **You should now have 25 binary features.**

Train-test split

6. We split the data into training and test sets with 80% of the data in the training set and 20% of the data in the test set. We use seed=1 so that everyone gets the same result. Using SFrame, this would look like:

```
train_data, test_data = loans_data.random_split(.8, seed=1)
```

(with **seed=1** to ensure people get the same results.)

If you are not using SFrame, download the list of indices for the training and test sets: module-8-assignment-2-train-idx.json (https://eventing.coursera.org/api/redirectStrict/pw-kEGNnvbTNamquCwanVAqL0XF8Ai37H0PWg1Zbla_P3M3VKUE40cf9B-2y9-hvh4gv69eFfQCukeTI60Vq_Q.HVQYDXOiWtRT_XfJDu7CSw.w2VonRsdEqBSWO1dSSl5KkBWLt1bGU8wf49X3smb8v1xlhV-Xz6Yg17irWoUxIMs6DuuVbFmfxzE9iTmP8LMBTm6v9b3ufLvk65ICSDPIlE9GpDdpq4j3YZcLUS9xss-TRPCR26rDik_ureuGYkR8YpyP-K9sP1yJ0N9Obs0BjJSn0zUi3o1wIgrMbqiDRNarPdJuUZQPRGOo_fbnNMhUvTipyv9AKtKA4ZbZwiCI1KKrHQrL0r9MgwG5ZAlf0BwknR98hdJr-2rRdVHxg_R-j4WlbNYeND0wZDN5_KHUmhXXYvRD1eMqyROJKri9XldUFtmv_OqUoplrGHIUU7dD15gxSwnQe5ufWbtkAWTJE6PqLHcArQQOceeXHdkb49wZpLceeBqTeCdJbd-Ndm7JvuJRoaMp7pUYg3JBgV_-oHoX5qVpHsNsvVkEfNU_7Qniy-lyNQZHK8uwGq4gXaKnA-wcCX1HKcNOGwfOi4rGwm6Et0gRisVfPo4-Ia1wl9B), module-8-assignment-2-test-idx.json (https://eventing.coursera.org/api/redirectStrict/xDNKCDHpW6PZNUsjS1Vud0hWs9yz9bvpYjYto9uvGGqgoEgv_e1ugk18K7x5wqGugUySY8M-Ux0lp3MX3zVRGA.c-YVSccYtzCRHcCepgWfbw.5KCSScbMrPW5sqCNOCKGjKnBfF-4oTyYqZVH0SnUZFzSNbJ11OOq3ANxh0sbelGUSLrX4iKgcv3DcYBPiTDwO19UM-zl3t9Af0oEfFnHHEkDyvKlH0zFvSsinvpL1XEK5-Izxp3YVoTUX_G9M_O6nTpyqZbCbWjioVasyvim0JWFeLxV3GMUNZN9Egu1su400-t6RqV7BLgWDb58LYL_PdENGlMtnynt7wQf5g77PILWUcd9YRaYa2DPSjXcaMCFWPfdRDNh0Am50WEJ71dhlS90fn86k4Vo9advi5QFageAVOzKiT-yYri5geo6tbvrVP0tLFXv3H6WdjMHj13iSgyi6rUXuipyJmInQ_bwQM7JwRVHb5b5cvhuw1x-Gcx_GNWr0dDqa6OJ3cKK7PBffqa2iCEw7GuUb12fb23EZ7Et8wnVFxRKq0o_vuVkTVXqWbDRRA96Bdk7AoKldCyrW4L-SxOjL8iSKWVznow8kPOQGDbqSkthlhGDixeRZQ7Z). IMPORTANT: If you are using a programming language with 1-based indexing (e.g. R, Matlab), make sure to increment all indices by 1.

Call the training and test sets **train_data** and **test_data**, respectively.

Weighted decision trees

7. Let's modify our decision tree code from Module 5 to support weighting of individual data points.

Weighted error definition

8. Consider a model with N data points with:

- Predictions $\hat{y}_1, ..., \hat{y}_n$
- Target $y_1, ..., y_n$
- Data point weights $\alpha_1, ..., \alpha_n$

Then the **weighted error** is defined by:

$$\sum^{n} \quad \cdot \quad 1[\quad / \quad]$$

where 1[y_i ≠ ŷ_i] is an indicator function that is set to 1 if y_i ≠ ŷ_i.

Write a function to compute weight of mistakes

9. Write a function that calculates the weight of mistakes for making the "weighted-majority" predictions for a dataset. The function accepts two inputs:

- **labels_in_node**: y_1, ..., y_n
- **data_weights**: Data point weights α_1, ..., α_n

We are interested in computing the (total) weight of mistakes, i.e.

$$\mathrm{WM}(\alpha, \hat{\mathbf{y}}) = \sum_{i=1}^{n} \alpha_i \times 1[y_i \neq \hat{y}_i].$$

This quantity is analogous to the number of mistakes, except that each mistake now carries different weight. It is related to the weighted error in the following way:

$$\mathrm{E}(\alpha, \hat{\mathbf{y}}) = \frac{\mathrm{WM}(\alpha, \hat{\mathbf{y}})}{\sum_{i=1}^{n} \alpha_i}$$

The function **intermediate_node_weighted_mistakes** should first compute two weights:

- WM(−1): weight of mistakes when all predictions are ŷ_i = −1 i.e. WM(**α**,**−1**)
- WM(+1): weight of mistakes when all predictions are ŷ_i = +1 i.e. WM(**α**,**+1**)

where **−1** and **+1** are vectors where all values are -1 and +1 respectively.

After computing WM(**−1**) and WM(**+1**), the function **intermediate_node_weighted_mistakes** should return the lower of the two weights of mistakes, along with the class associated with that weight. The function should be analogous to the following Python function:

```
def intermediate_node_weighted_mistakes(labels_in_node, data_weights):
    # Sum the weights of all entries with label +1
    total_weight_positive = sum(data_weights[labels_in_node == +1])

    # Weight of mistakes for predicting all -1's is equal to the sum above
    ### YOUR CODE HERE
    weighted_mistakes_all_negative = ...

    # Sum the weights of all entries with label -1
    ### YOUR CODE HERE
    total_weight_negative = ...

    # Weight of mistakes for predicting all +1's is equal to the sum above
    ### YOUR CODE HERE
    weighted_mistakes_all_positive = ...

    # Return the tuple (weight, class_label) representing the lower of the two weights
    #    class_label should be an integer of value +1 or -1.
    # If the two weights are identical, return (weighted_mistakes_all_positive,+1)
    ### YOUR CODE HERE
    ...
```

10. Recall that the **classification error** is defined as follows:

$$\text{classification error} = \frac{\#\text{ mistakes}}{\#\text{ all data points}}$$

Quiz Question: If we set the weights $\alpha=1$ for all data points, how is the weight of mistakes $WM(\alpha,\hat{y})$ related to the classification error?

Function to pick best feature to split on

11. We continue modifying our decision tree code from the earlier assignment to incorporate weighting of individual data points. The next step is to pick the best feature to split on.

The **best_splitting_feature** function is similar to the one from the earlier assignment with two minor modifications:

- The function **best_splitting_feature** should now accept an extra parameter data_weights to take account of weights of data points.
- Instead of computing the number of mistakes in the left and right side of the split, we compute the weight of mistakes for both sides, add up the two weights, and divide it by the total weight of the data.

Your function should be analogous to the following Python function:

```
# If the data is identical in each feature, this function should return None

def best_splitting_feature(data, features, target, data_weights):

    # These variables will keep track of the best feature and the corresponding error
    best_feature = None
    best_error = float('+inf')
    num_points = float(len(data))

    # Loop through each feature to consider splitting on that feature
    for feature in features:

        # The left split will have all data points where the feature value is 0
        # The right split will have all data points where the feature value is 1
        left_split = data[data[feature] == 0]
        right_split = data[data[feature] == 1]

        # Apply the same filtering to data_weights to create left_data_weights, right_data_weights
        ## YOUR CODE HERE
        left_data_weights = ...
        right_data_weights = ...

        # DIFFERENT HERE
        # Calculate the weight of mistakes for left and right sides
        ## YOUR CODE HERE
        left_weighted_mistakes, left_class = ...
        right_weighted_mistakes, right_class = ...

        # DIFFERENT HERE
        # Compute weighted error by computing
        #  ( [weight of mistakes (left)] + [weight of mistakes (right)] ) / [total weight of all data
points]
        ## YOUR CODE HERE
        error = ...

        # If this is the best error we have found so far, store the feature and the error
        if error < best_error:
            best_feature = feature
            best_error = error

    # Return the best feature we found
    return best_feature
```

Very Optional. Relationship between weighted error and weight of mistakes

By definition, the weighted error is the weight of mistakes divided by the weight of all data points, so

$$\mathrm{E}(\alpha, \hat{\mathbf{y}}) = \frac{\sum_{i=1}^{n} \alpha_i \times 1[y_i \neq \hat{y}_i]}{\sum_{i=1}^{n} \alpha_i} = \frac{\mathrm{WM}(\alpha, \hat{\mathbf{y}})}{\sum_{i=1}^{n} \alpha_i}$$

In the code above, we obtain E(α, ŷ) from the two weights of mistakes from both sides, WM(α_left, ŷ_left) and WM(α_right, ŷ_right). First, notice that the overall weight of mistakes WM(α, ŷ) can be broken into two weights of mistakes over either side of the split:

$$\mathrm{WM}(\alpha, \hat{\mathbf{y}}) = \sum_{i=1}^{n} \alpha_i \times 1[y_i \neq \hat{y}_i] = \sum_{\text{left}} \alpha_i \times 1[y_i \neq \hat{y}_i] + \sum_{\text{right}} \alpha_i \times 1[y_i \neq \hat{y}_i]$$

We then divide through by the total weight of all data points to obtain E(α, ŷ):

Building the tree

12. With the above functions implemented correctly, we are now ready to build our decision tree. Recall from the previous assignments that each node in the decision tree is represented as a dictionary which contains the following keys:

```
{
   'is_leaf'            : True/False.
   'prediction'         : Prediction at the leaf node.
   'left'               : (dictionary corresponding to the left tree).
   'right'              : (dictionary corresponding to the right tree).
   'features_remaining' : List of features that are posible splits.
}
```

Let us start with a function that creates a leaf node given a set of target values. The **create_leaf** function should be analogous to the following cell:

```
def create_leaf(target_values, data_weights):

    # Create a leaf node
    leaf = {'splitting_feature' : None,
            'is_leaf': True}

    # Computed weight of mistakes.
    # Store the predicted class (1 or -1) in leaf['prediction']
    weighted_error, best_class = intermediate_node_weighted_mistakes(target_values, data_weights)
    leaf['prediction'] = ... ## YOUR CODE HERE

    return leaf
```

13. Now write a function that learns a weighted decision tree recursively and implements 3 stopping conditions:

- All data points in a node are from the same class.
- No more features to split on.
- Stop growing the tree when the tree depth reaches **max_depth**.

Since there are many steps involved, we provide you with a Python skeleton, along with explanatory comments.

```
def weighted_decision_tree_create(data, features, target, data_weights, current_depth = 1, max_depth
= 10):
    remaining_features = features[:] # Make a copy of the features.
    target_values = data[target]
    print "--------------------------------------------------------------------"
    print "Subtree, depth = %s (%s data points)." % (current_depth, len(target_values))

    # Stopping condition 1. Error is 0.
    if intermediate_node_weighted_mistakes(target_values, data_weights)[0] <= 1e-15:
        print "Stopping condition 1 reached."
        return create_leaf(target_values, data_weights)

    # Stopping condition 2. No more features.
    if remaining_features == []:
        print "Stopping condition 2 reached."
        return create_leaf(target_values, data_weights)

    # Additional stopping condition (limit tree depth)
    if current_depth > max_depth:
        print "Reached maximum depth. Stopping for now."
        return create_leaf(target_values, data_weights)

    # If all the datapoints are the same, splitting_feature will be None. Create a leaf
    splitting_feature = best_splitting_feature(data, features, target, data_weights)
    remaining_features.remove(splitting_feature)

    left_split = data[data[splitting_feature] == 0]
    right_split = data[data[splitting_feature] == 1]

    left_data_weights = data_weights[data[splitting_feature] == 0]
    right_data_weights = data_weights[data[splitting_feature] == 1]

    print "Split on feature %s. (%s, %s)" % (\
              splitting_feature, len(left_split), len(right_split))

    # Create a leaf node if the split is "perfect"
    if len(left_split) == len(data):
        print "Creating leaf node."
        return create_leaf(left_split[target], data_weights)
    if len(right_split) == len(data):
        print "Creating leaf node."
        return create_leaf(right_split[target], data_weights)

    # Repeat (recurse) on left and right subtrees
    left_tree = weighted_decision_tree_create(
        left_split, remaining_features, target, left_data_weights, current_depth + 1, max_depth)
    right_tree = weighted_decision_tree_create(
        right_split, remaining_features, target, right_data_weights, current_depth + 1, max_depth)

    return {'is_leaf'          : False,
            'prediction'       : None,
            'splitting_feature': splitting_feature,
            'left'             : left_tree,
            'right'            : right_tree}
```

14. Finally, write a recursive function to count the nodes in your tree. The function should be analogous to

```
def count_nodes(tree):
    if tree['is_leaf']:
        return 1
    return 1 + count_nodes(tree['left']) + count_nodes(tree['right'])
```

Making predictions with a weighted decision tree

15. To make a single prediction, we must start at the root and traverse down the decision tree in recursive fashion. Write a function **classify** that makes a single prediction. It should be analogous to the following:

```
def classify(tree, x, annotate = False):
    # If the node is a leaf node.
    if tree['is_leaf']:
        if annotate:
            print "At leaf, predicting %s" % tree['prediction']
        return tree['prediction']
    else:
        # Split on feature.
        split_feature_value = x[tree['splitting_feature']]
        if annotate:
            print "Split on %s = %s" % (tree['splitting_feature'], split_feature_value)
        if split_feature_value == 0:
            return classify(tree['left'], x, annotate)
        else:
            return classify(tree['right'], x, annotate)
```

Evaluating the tree

16. Create a function called **evaluate_classification_error**. It takes in as input:

- **tree** (as described above)
- **data** (an data frame)

The function does not change because of adding data point weights. It is analogous to this Python function:

```
def evaluate_classification_error(tree, data):
    # Apply the classify(tree, x) to each row in your data
    prediction = data.apply(lambda x: classify(tree, x))

    # Once you've made the predictions, calculate the classification error
    return (prediction != data[target]).sum() / float(len(data))
```

Example: Training a weighted decision tree

17. To build intuition on how weighted data points affect the tree being built, consider the following:

Suppose we only care about making good predictions for the **first 10 and last 10 items** in train_data, we assign weights:

- 1 to the last 10 items
- 1 to the first 10 items
- and 0 to the rest.

Let us fit a weighted decision tree with max_depth = 2. Then compute the classification error on the **subset_20**, i.e. the subset of data points whose weight is 1 (namely the first and last 10 data points).

```
# Assign weights
example_data_weights = graphlab.SArray([1.] * 10 + [0.]*(len(train_data) - 20) + [1.] * 10)
# Train a weighted decision tree model.
small_data_decision_tree_subset_20 = weighted_decision_tree_create(train_data, features, target,
                         example_data_weights, max_depth=2)
```

18. Now, we will compute the classification error on the **subset_20**, i.e. the subset of data points whose weight is 1 (namely the first and last 10 data points).

```
evaluate_classification_error(small_data_decision_tree_subset_20, train_data)
```

The model small_data_decision_tree_subset_20 performs **a lot** better on subset_20 than on train_data.

So, what does this mean?

- The points with higher weights are the ones that are more important during the training process of the weighted decision tree.
- The points with zero weights are basically ignored during training.

Quiz Question: Will you get the same model as **small_data_decision_tree_subset_20** if you trained a decision tree with only the 20 data points with non-zero weights from the set of points in **subset_20**?

Implementing your own Adaboost (on decision stumps)

19. Now that we have a weighted decision tree working, it takes only a bit of work to implement Adaboost. For the sake of simplicity, let us stick with decision tree stumps by training trees with max_depth=1.

Recall from the lecture the procedure for Adaboost:

* Start with unweighted data with $\alpha_j = 1$

* For t=1,...,T:

- Learn $f_t(x)$ with data weights α_j
- Compute coefficient $\hat{w}_t$:

- Re-compute weights α_j

- Normalize weights α_j

Now write your own Adaboost function. The function accepts 4 parameters:

- data: a data frame with binary features
- features: list of feature names
- target: name of target column
- num_tree_stumps: number of tree stumps to train for the ensemble

The function should return the list of tree stumps, along with the list of corresponding tree stump weights.

It should be analogous to the following code skeleton:

```
from math import log
from math import exp

def adaboost_with_tree_stumps(data, features, target, num_tree_stumps):
    # start with unweighted data
    alpha = graphlab.SArray([1.]*len(data))
    weights = []
    tree_stumps = []
    target_values = data[target]

    for t in xrange(num_tree_stumps):
        print '====================================================='
        print 'Adaboost Iteration %d' % t
        print '====================================================='
        # Learn a weighted decision tree stump. Use max_depth=1
        tree_stump = weighted_decision_tree_create(data, features, target, data_weights=alpha, max_de
pth=1)
        tree_stumps.append(tree_stump)

        # Make predictions
        predictions = data.apply(lambda x: classify(tree_stump, x))

        # Produce a Boolean array indicating whether
        # each data point was correctly classified
        is_correct = predictions == target_values
        is_wrong   = predictions != target_values

        # Compute weighted error
        # YOUR CODE HERE
        weighted_error = ...

        # Compute model coefficient using weighted error
        # YOUR CODE HERE
        weight = ...
        weights.append(weight)

        # Adjust weights on data point
        adjustment = is_correct.apply(lambda is_correct : exp(-weight) if is_correct else exp(weigh
t))

        # Scale alpha by multiplying by adjustment
        # Then normalize data points weights
        ## YOUR CODE HERE
        ...

    return weights, tree_stumps
```

Reminders

- Stump weights ($\hat{w}$) and data point weights (α) are two different concepts.
- Stump weights ($\hat{w}$) tell you how important each stump is while making predictions with the entire boosted ensemble.
- Data point weights (α) tell you how important each data point is while training a decision stump.

Training a boosted ensemble of 10 stumps

20. Let us train an ensemble of 10 decision tree stumps with Adaboost. We run the **adaboost_with_tree_stumps** function with the following parameters:

- train_data
- features

- target
- num_tree_stumps = 10

Making predictions

21. Recall from the lecture that in order to make predictions, we use the following formula:

Do the following things in a new function **predict_adaboost**:

- Compute the predictions f_t(x) using the t-th decision tree
- Compute ŵ_t f_t(x) by multiplying the stump_weights with the predictions f_t(x) from the decision trees
- Sum the weighted predictions over each stump in the ensemble.

In the end, your **predict_adaboost** should be analogous to this Python function:

```
def predict_adaboost(stump_weights, tree_stumps, data):
    scores = graphlab.SArray([0.]*len(data))

    for i, tree_stump in enumerate(tree_stumps):
        predictions = data.apply(lambda x: classify(tree_stump, x))

        # Accumulate predictions on scaores array
        # YOUR CODE HERE
        ...

    return scores.apply(lambda score : +1 if score > 0 else -1)
```

Use this function to answer the following question:

Quiz Question: Are the weights monotonically decreasing, monotonically increasing, or neither?

Reminder: Stump weights (ŵ) tell you how important each stump is while making predictions with the entire boosted ensemble.

Performance plots

How does accuracy change with adding stumps to the ensemble?

22. We will now train an ensemble with:

- train_data
- features
- target
- num_tree_stumps = 30

Once we are done with this, we will then do the following:

- Compute the classification error at the end of each iteration.
- Plot a curve of classification error vs iteration.

First, let's train the model.

Computing training error at the end of each iteration

23. Let us compute the classification error on the **train_data** and see how it is reduced as trees are added.

For n = 1 to 30, do the following:

- Make predictions on **train_data** using tree stumps 0, ..., n-1.
- Compute classification error for the predictions
- Record the classification error for that n.

The loop should be analogous to the following:

```
error_all = []
for n in xrange(1, 31):
    predictions = predict_adaboost(stump_weights[:n], tree_stumps[:n], train_data)
    error = 1.0 - graphlab.evaluation.accuracy(train_data[target], predictions)
    error_all.append(error)
    print "Iteration %s, training error = %s" % (n, error_all[n-1])
```

Visualizing training error vs number of iterations

24. Let us generate the plot of classification error as a function of the number of iterations. Use the classification error values recorded in #23.

For inspiration, we provide you with matplotlib plotting code.

```
plt.rcParams['figure.figsize'] = 7, 5
plt.plot(range(1,31), error_all, '-', linewidth=4.0, label='Training error')
plt.title('Performance of Adaboost ensemble')
plt.xlabel('# of iterations')
plt.ylabel('Classification error')
plt.legend(loc='best', prop={'size':15})

plt.rcParams.update({'font.size': 16})
```

Quiz Question: Which of the following best describes a **general trend in accuracy** as we add more and more components? Answer based on the 30 components learned so far.

- Training error goes down monotonically, i.e. the training error reduces with each iteration but never increases.
- Training error goes down in general, with some ups and downs in the middle.
- Training error goes up in general, with some ups and downs in the middle.
- Training error goes down in the beginning, achieves the best error, and then goes up sharply.
- None of the above

Evaluation on the test data

25. Performing well on the training data is cheating, so lets make sure it works on the test_data as well. Here, we will compute the classification error on the test_data at the end of each iteration.

For n = 1 to 30, do the following:

- Make predictions on **test_data** using tree stumps 0, ..., n-1.
- Compute classification error for the predictions

- Record the classification error for that n.

Visualize both the training and test errors

26. Let us plot the training & test error with the number of iterations.

Again, for inspiration, we provide you with matplotlib code.

```
plt.rcParams['figure.figsize'] = 7, 5
plt.plot(range(1,31), error_all, '-', linewidth=4.0, label='Training error')
plt.plot(range(1,31), test_error_all, '-', linewidth=4.0, label='Test error')

plt.title('Performance of Adaboost ensemble')
plt.xlabel('# of iterations')
plt.ylabel('Classification error')
plt.rcParams.update({'font.size': 16})
plt.legend(loc='best', prop={'size':15})
plt.tight_layout()
```

Quiz Question: From this plot (with 30 trees), is there massive overfitting as the # of iterations increases?